

Article

Surrogate Modeling of Methane Yield in Anaerobic Digestion Under Heterogeneous Feedstock Conditions

Kamil Witaszek ¹, Nikolay Kiktev ², Volodymyr Reshettiuk ³ and Tomasz Garbowski ^{1,*}

¹ Department of Biosystems Engineering, Poznan University of Life Sciences, Wojska Polskiego 50, 60-627 Poznań, Poland; kamil.witaszek@up.poznan.pl

² Department of Automation and Robotic Systems named after Acad. I.I. Martynenko, National University of Life and Environmental Sciences of Ukraine, Heroiv Oborony Str., 03041 Kyiv, Ukraine; nkiktev@ukr.net

³ Department of Fundamentals of Engineering and Power Engineering, Warsaw University of Life Sciences (SGGW), 02-787 Warsaw, Poland; volodymyr_reshettiuk@sggw.edu.pl

* Correspondence: tomasz.garbowski@up.poznan.pl

Abstract

Predicting methane and biogas yields from heterogeneous feedstocks requires distinguishing performance within an available dataset from transfer to unrepresented substrate labels. This study compares artificial neural network (ANN), Gaussian process regression (GP), and proper orthogonal decomposition combined with radial basis function (POD–RBF) using 83 laboratory observations and five measured descriptors. The two measured fresh-mass yields are the primary endpoints; four DM/ODM-normalized yields are secondary. Shared five-fold cross-validation (CV) and 30-label leave-one-substrate-out (LOSO) validation exclude substrate identity from the predictors and fit preprocessing and POD exclusively on training data. For the primary endpoints, POD–RBF has the lowest mean CV root mean square error (23.84 m³/Mg fresh mass), whereas GP has the lowest mean CV symmetric mean absolute percentage error (sMAPE; 29.71%). Mean group-wise LOSO sMAPE is 46.35%, 46.78%, and 45.09% for ANN, GP, and POD–RBF, respectively. Mean coverage of conditional nominal 95% GP observation intervals decreases from 93.37% in CV to 77.11% in LOSO. Three full-data POD modes retain 99.16% of centered output energy, partly reflecting deterministic relationships among mass-normalized yields. Model rankings depend on the metric and validation setting, and the single CV partition leaves split dependence unquantified. Large errors for some excluded labels and GP under-coverage limit claims of reliable prediction for unseen feedstocks or industrial application.

Keywords: anaerobic digestion; methane yield; biogas; surrogate modeling; Gaussian Process; proper orthogonal decomposition; grouped validation



Academic Editor: Prasad Kaparaju

Received: 4 September 2026

Revised: 4 October 2026

Accepted: 7 October 2026

Published: 12 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Anaerobic digestion converts organic residues into methane-rich biogas and supports energy recovery from agricultural and other organic wastes. Feedstock preparation and operating conditions influence the resulting yields, making prediction relevant to experimental planning and substrate assessment [1,2]. Demand-driven biogas production further increases the value of reliable estimates of available energy [3]. However, feedstocks differ in composition and biodegradability: experiments on complex organic substrates show that theoretical methane potential can overestimate measured production when the degradable fraction is not accounted for [4]. Przybył et al. examined the effect of ensiling enhancers on

the methane yield of corn stover, further linking substrate preparation to its performance in anaerobic digestion [5].

The experimental definition of yield also matters. Biochemical methane potential (BMP) protocols specify assay conditions and reporting conventions, while efforts toward standardization address variability associated with inoculum, substrate preparation, test setup and data treatment [6,7]. Harmonizing column names and units is necessary when pooling records, but does not establish equivalence of the underlying experiments. Assay duration, gas-reference conditions and normalization basis can affect yield comparisons, making data provenance essential to model interpretation.

Mechanistic and data-driven approaches address this prediction problem at different levels of detail. The Anaerobic Digestion Model No. 1 (ADM1) represents biochemical conversion steps together with physicochemical processes [8]. Recent full-scale comparisons of a simplified ADM1 with random forests and recurrent neural networks illustrate trade-offs among predictive performance, input requirements and computational effort [9]. A surrogate fitted to laboratory observations instead estimates an empirical descriptor-to-yield mapping without resolving reaction dynamics.

Neural-network applications to anaerobic digestion have a substantial history, including hierarchical networks for process control [10]. Subsequent studies have addressed full-scale production monitoring, deep-learning prediction of co-digestion and the usefulness of available input variables [11–13]. Other applications estimate total biogas potential from early test outputs or silage energy value from measured characteristics [14,15]. Recent work also includes optimized multilayer perceptrons, household-waste prediction and stacking ensembles [16–18]. These applications differ in both their prediction targets and the information available at prediction time.

Long et al. used several years of commercial-digester data spanning diverse feedstocks and operating conditions [19], whereas Meola and Weinrich investigated dynamic methane production at intra-day resolution [20]. Such temporal forecasting tasks use a different information structure from predicting laboratory yields for heterogeneous substrates. Comparison of surrogate models therefore requires common observations, predictors, output definitions and validation partitions.

Validation must also match the intended prediction setting. Random partitions can place observations from the same substrate label in both training and test sets; holding out that label asks a different question about transfer between groups. Studies of structured data show why partitioning changes the interpretation of predictive error [21]. Small samples further make performance estimates uncertain [22], and preprocessing or model selection that uses test information can produce optimistic results [23]. This motivates training-only transformations and comparison of random-row and substrate-label holdout performance, alongside recognition of the need for external validation.

ANN, Gaussian process regression (GP), and proper orthogonal decomposition with radial basis function (POD–RBF) provide complementary representations. A shared-hidden-layer ANN couples multiple outputs; GP specifies a probabilistic regression model through its covariance structure [24]; POD–RBF learns coefficients in a reduced output space. POD-based models have been used for chemical monitoring and engineering inverse analysis [25,26]. Garbowski et al. also combined dimensional reduction and surrogate approximation in the calibration of orthotropic material models [27]. These precedents motivate testing output compression for heterogeneous digestion data.

Predictive uncertainty requires separate evaluation. The availability of a probabilistic model does not ensure that nominal intervals achieve their stated coverage; regression-calibration research makes this distinction explicit [28]. Within anaerobic digestion, recent work has combined grouped cross-study validation, ensemble learning and conformal

uncertainty estimation on a substantially larger literature-derived dataset [29]. Here, empirical coverage under both validation protocols tests the reliability of the GP intervals.

A further issue concerns the response representation. Methane and biogas yields are frequently reported per unit of fresh mass, dry matter or organic dry matter. In the present workbook, four of the six outputs are deterministic conversions of the two fresh-mass yields using mass fractions that also enter as predictors. Consequently, a compact POD basis can reflect reporting identities as well as empirical covariation. Reconstruction accuracy therefore cannot establish an independent multi-output advantage; predictions must be assessed on held-out observations.

Against this background, the contribution is a reproducible comparison of three prescribed surrogate formulations on the same 83 experimental observations. It examines whether model rankings change across error metrics and validation settings, whether nominal GP intervals retain coverage under substrate-label holdout, and how deterministic output dependence affects interpretation of POD compression. The aim is to clarify what these models can support with a small heterogeneous dataset; it is not to introduce a new learning algorithm or establish universal superiority of one model family.

Five continuous physicochemical descriptors are used as predictors, while substrate labels define the leave-one-substrate-out (LOSO) partitions. All predictive preprocessing and POD fitting are confined to training data. Auxiliary relevance and sensitivity analyses complement the comparison, and one documented computational workflow links the source records to the reported tables and figures.

2. Materials and Methods

2.1. Dataset, Variables, and Problem Formulation

The dataset comprises 83 experimental observations obtained entirely from studies conducted by the authors and their research teams. A substantial part originated from laboratory experiments at the Department of Biosystems Engineering, Poznan University of Life Sciences; the remaining observations originated from other experimental studies previously conducted by the co-authors. The dataset was not assembled from unrelated external databases or literature sources.

Measurements were obtained in laboratory-scale anaerobic digestion experiments with different feedstocks. According to the contributing research teams, comparable experimental and analytical procedures were used to determine basic feedstock characteristics and biogas/methane production. This account is attributed to those teams and does not constitute independent verification of protocol equivalence. The records were compiled into common variables and units.

The input vector contains five continuous feedstock and process descriptors:

$$x = [PH, COND, TIME, DM, ODM] \quad (1)$$

where:

- SILA is retained separately as the literal substrate label used to construct the LOSO groups. It is not a numerical predictor in Equation (1).
- PH denotes the measured acidity level.
- COND corresponds to electrical conductivity.
- TIME denotes the recorded fermentation duration in days; its use assumes this duration is known for the prediction task.
- DM is dry matter expressed as a percentage of fresh mass.
- ODM is organic dry matter expressed as a percentage of dry matter.

The predictive task consists of estimating methane and biogas production expressed using several normalization schemes. The output vector is therefore defined as

$$\mathbf{y} = [\text{CH4}_{\text{fm}}, \text{Biogas}_{\text{fm}}, \text{CH4}_{\text{dm}}, \text{Biogas}_{\text{dm}}, \text{CH4}_{\text{odm}}, \text{Biogas}_{\text{odm}}] \quad (2)$$

The subscripts fm, dm and odm indicate fresh mass, dry matter and organic dry matter. Gas yield is expressed in m^3/Mg ; one Mg equals one metric tonne. The four normalized outputs are calculated from the two fresh-mass yields and the measured mass fractions.

From a modeling perspective, the problem can be expressed as nonlinear multi-output regression

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon, \quad (3)$$

where $f(\mathbf{x})$ represents the unknown nonlinear mapping between substrate characteristics and methane production, while ε denotes measurement noise and model discrepancy.

The compiled dataset does not retain experiment-level metadata on reactor configuration, inoculum, temperature, gas-volume reference conditions or blank correction. Comparable procedures are reported by the data contributors, but equivalence of all experimental protocols cannot be independently verified from the compiled table. The supplied fresh-mass yields and units are retained without undocumented laboratory or gas-reference corrections. This limits interpretation across experimental sources.

In the present work three complementary surrogate modeling strategies are investigated: (I) deterministic nonlinear regression using artificial neural networks (ANNs); (II) probabilistic regression using Gaussian processes (GPs); (III) reduced-order multi-output modeling using proper orthogonal decomposition combined with radial basis function regression (POD–RBF).

Comparative performance is evaluated using the protocols in Section 2.4. The source workbook and row-preserving CSV are available from the corresponding author upon request.

2.2. Data Preprocessing and Feature Representation

Table 1 summarizes the experimental workbook. There are 83 complete numeric records and 30 distinct literal SILA labels: 26 labels have three observations, Whey II has two, and Sugar beet pulp W1, W2 and W3 have one each. The three W-labeled records are retained as separate labels exactly as supplied; no biological distinction or common-batch identity is inferred from their names.

Table 1. Descriptive statistics calculated from the 83 records. SD is the sample standard deviation ($n - 1$); values are rounded to two decimal places for presentation. SILA is used only for grouping, not as a predictor.

Variable	Unit	Role	Range or Labels	Mean $\pm$ SD or Count
SILA	–	Grouping metadata	30 literal labels	83 records
PH	–	Input	3.50–9.13	7.05 ± 1.57
COND	mS/cm	Input	0.54–26.92	14.68 ± 9.21
TIME	days	Input	15.00–63.00	28.80 ± 10.66
DM	% FM	Input	1.82–90.51	26.94 ± 24.40
ODM	% DM	Input	53.01–98.24	87.16 ± 11.29
CH4_fm	$\text{m}^3/\text{Mg FM}$	Output	0.91–207.88	72.83 ± 59.92
Biogas_fm	$\text{m}^3/\text{Mg FM}$	Output	2.37–413.84	129.29 ± 103.71
CH4_dm	$\text{m}^3/\text{Mg DM}$	Output	27.23–790.46	280.89 ± 137.47
Biogas_dm	$\text{m}^3/\text{Mg DM}$	Output	71.58–1123.95	504.59 ± 219.98
CH4_odm	$\text{m}^3/\text{Mg ODM}$	Output	46.95–925.94	317.74 ± 151.10
Biogas_odm	$\text{m}^3/\text{Mg ODM}$	Output	127.29–1321.84	571.04 ± 232.24

All 83 records are retained in the dataset summaries and predictive validation. The source sheet, row identifiers and literal labels are preserved in the computational files, providing a direct mapping between the workbook and every held-out prediction. No small group is omitted or combined with another label.

2.2.1. Substrate Labels and Predictor Representation

SILA is categorical. Feeding integer identifiers to a distance-based GP or RBF kernel would impose artificial distances between unrelated labels, regardless of whether the authors intend an ordinal interpretation. A common numeric code can also be exploited by an ANN and does not become a physical descriptor simply because all models receive the same input.

For grouping, the observed labels form a finite set:

$$SILA \in G, \quad |G| = 30, \quad (4)$$

where G is the set of the 30 literal labels in the source workbook. Numeric keys in exported files index this set for storage only.

All principal models therefore use PH, COND, TIME, DM and ODM alone. SILA is excluded from prediction, ARD, permutation importance and Sobol analysis. One-hot coding would remove an imposed ordinal distance for observed labels, but would still provide no learned effect for a previously unseen label. Descriptor-only inputs directly match the intended label-holdout task.

LOSO assesses transfer between the supplied labels, not necessarily between independent biological substrate classes or experimental campaigns. Several labels name related materials, and three labels contain a single observation. These characteristics are reported explicitly and limit how far label-holdout results can be generalized to genuinely new industrial feedstocks.

2.2.2. Standardization of Continuous Variables

The remaining input variables (PH, COND, TIME, DM, ODM) represent continuous quantities with different physical units and ranges. Such scale differences can negatively affect training stability and convergence, particularly in kernel-based models and gradient-driven optimization procedures. Variables with larger numeric ranges may dominate distance computations or gradient magnitudes, biasing model fitting toward specific predictors for purely numerical reasons.

To mitigate this effect, continuous variables are standardized using z-score normalization:

$$x_i^{\text{norm}} = \frac{x_i - \mu_i}{\sigma_i}, \quad (5)$$

where μ_i and σ_i denote the mean and standard deviation computed from the training dataset.

This transformation shifts each variable to zero mean and rescales it to unit variance, thereby improving numerical conditioning and enabling fair comparison of feature influence across variables. Importantly, the normalization parameters μ_i and σ_i are computed exclusively using training data within each validation split, and then applied to the corresponding validation data. This split-specific normalization ensures that no information from the validation subsets is inadvertently used during preprocessing, preventing leakage that could lead to overly optimistic performance estimates.

Continuous inputs are standardized using their training means and population standard deviations. Each output is similarly standardized for ANN and GP. POD instead centers the unscaled outputs. A zero training standard deviation is replaced by one. The

same fitted transformation is applied to the held-out rows, with no preprocessing fitted to the combined training and test set.

2.2.3. Multi-Output Structure

The six outputs are mathematically dependent. For either methane or biogas, dry-matter yield is calculated by dividing fresh-mass yield by DM/100, and organic-dry-matter yield by subsequently dividing by ODM/100. The two measured fresh-mass outputs are the primary performance endpoints; the four normalized outputs are secondary calculated quantities. This distinction does not imply that methane and total biogas yields are statistically independent, because they are related through gas composition.

Consequently, a high POD energy concentration can partly reflect normalization identities already involving two model inputs. It cannot by itself demonstrate an independent multi-output learning advantage. Sharing hidden units or restricting predictions to a POD subspace also does not enforce the row-specific conversion identities exactly. Figure 1 summarizes the three predictive representations.

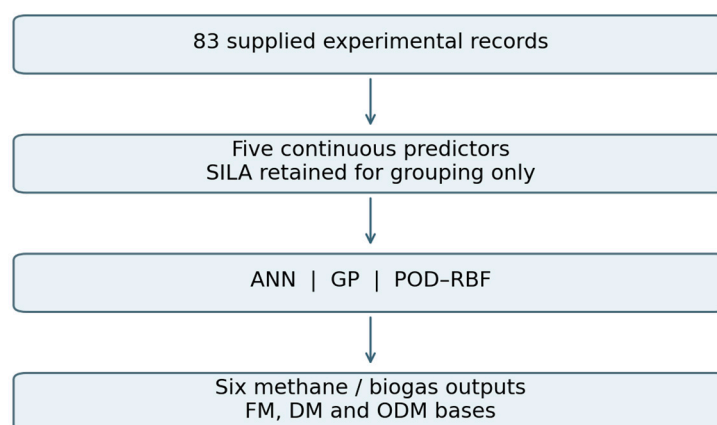


Figure 1. Structure of the prediction problem. Five continuous descriptors predict six yield quantities; SILA is used for grouping only.

GP is fitted separately for each output, while ANN shares a hidden representation and POD-RBF predicts modal coefficients. All models retain their six-output formulation. Primary comparisons use only methane-FM and biogas-FM performance, while secondary normalized-output results describe the remaining targets. The comparison does not isolate the benefit of output coupling from deterministic mass conversion.

2.3. Performance Metrics for Multi-Output Regression

Prediction accuracy is evaluated using several complementary metrics that capture different aspects of surrogate model performance. Because the problem involves nonlinear regression with multiple outputs and heterogeneous data ranges, both absolute-error and relative-error measures are reported. Using multiple metrics allows the predictive performance of the models to be assessed more robustly and reduces the risk that conclusions depend on a single evaluation criterion.

The first measure considered is the root mean square error (RMSE), which quantifies the magnitude of prediction errors in the same physical units as the modeled variable. RMSE penalizes larger deviations more strongly due to the quadratic error term and therefore highlights cases in which the model produces substantial prediction errors. It is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (6)$$

where y_i denotes the measured value and $\hat{y}_i$ the corresponding model prediction.

While RMSE provides information about the absolute magnitude of prediction errors, it does not describe how well the model explains the variability of the data. For this reason, the coefficient of determination R^2 is also reported. This metric evaluates the fraction of variance in the observed data that can be explained by the model predictions. It is computed as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (7)$$

Here the reference mean is calculated over the held-out observations used for the reported metric. Negative R^2 indicates performance worse than their constant mean. R^2 is reported for pooled CV outputs rather than individual small LOSO groups; it is undefined if the reference sum of squares is numerically zero.

Although RMSE and R^2 are widely used measures of predictive accuracy, they can be difficult to interpret when output variables differ substantially in magnitude or units. This is particularly relevant in anaerobic digestion studies, where methane and biogas yields may vary significantly across substrates and normalization schemes. To address this issue, relative-error metrics expressed as percentages are also considered.

The mean absolute percentage error (MAPE) quantifies the average relative deviation between predictions and measurements:

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{|y_i|} \right|. \quad (8)$$

MAPE uses the absolute measured value in the denominator and no added epsilon. It is undefined if any measured value is zero. All supplied reference yields are positive, so no observations are omitted from the present MAPE calculations.

However, MAPE may become unstable when the true value y_i approaches zero, because the denominator becomes very small. This issue can lead to disproportionately large percentage errors even when the absolute prediction error is moderate. To mitigate this effect, the symmetric mean absolute percentage error (sMAPE) is also employed:

$$\text{sMAPE} = \frac{100}{n} \sum_{i=1}^n \frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}. \quad (9)$$

sMAPE is bounded between 0% and 200%, including when a model predicts a negative yield. A pair of zero measured and predicted values contributes zero. Predictions are not clipped to non-negative values; negative estimates count as errors and reveal a limitation of these unconstrained regression models.

2.4. Validation Protocols: K-Fold Cross-Validation and Leave-One-Substrate-Out (LOSO)

Reliable evaluation of surrogate models requires validation strategies that capture both interpolation accuracy within the observed dataset and the ability of the models to generalize beyond the conditions represented in the training data. In the context of anaerobic digestion modeling, this distinction is particularly important because experimental datasets often contain relatively small numbers of observations distributed across heterogeneous substrate groups.

Five-fold random-row CV and 30-label LOSO use identical partitions across model families. The former assesses prediction among the available records; the latter tests

omission of an entire supplied label. Neither protocol constitutes prospective or external industrial validation.

2.4.1. K-Fold Cross-Validation

Formally, the dataset D containing all observations is partitioned into K disjoint subsets:

$$D = \cup_{k=1}^K D_k. \quad (10)$$

Each subset D_k serves once as a validation set while the remaining $K - 1$ subsets are used for model training. For a given fold k , the training dataset is therefore defined as

$$D_{\text{train}}^{(k)} = D \setminus D_k, \quad (11)$$

while D_k represents the corresponding validation dataset.

Every fold is fitted using its own training records, including input scaling, output scaling or centering, POD decomposition and retained rank. No held-out row contributes to these quantities. The fixed main model settings are specified before this run, rather than selected using outer-fold errors. Training and test row identifiers are exported for every fit.

2.4.2. Leave-One-Substrate-Out Validation

While k-fold cross-validation evaluates model performance within the distribution represented by the dataset, it does not fully capture the challenge of predicting digestion performance for new substrates not present in the training data. In practical applications, surrogate models may be used to estimate methane or biogas production for feedstocks that were not included in the original experimental dataset.

LOSO excludes all records with one literal SILA label, trains on the remaining records and predicts the excluded records. This is repeated for all 30 labels, including the three singleton W labels. Continuous descriptors alone are available to the predictor.

Let g denote a substrate group within the set of all substrate categories. The LOSO training dataset is defined as

$$D_{\text{train}}^{(g)} = D \setminus D_g, \quad (12)$$

while the corresponding test dataset is

$$D_{\text{test}}^{(g)} = D_g, \quad (13)$$

where D_g denotes all observations belonging to substrate group g .

The test label is unseen during fitting, but related materials or experiments may remain in training under other labels. The compiled data do not identify experimental campaigns independently of SILA. Thus, the protocol tests transfer between recorded labels and should not be interpreted as a verified leave-one-independent-experimental-source-out design.

The group-wise errors reveal where descriptor-based prediction deteriorates when a label is absent from training. These errors are interpreted together with the small and unequal group sizes and the incomplete information on experimental provenance at row level.

LOSO errors are summarized by label, ranked within each model, and compared through a lowest-error model map. These displays use the same group sMAPE values and describe this dataset rather than universal generalization ability. Figure 2 illustrates both validation schemes.

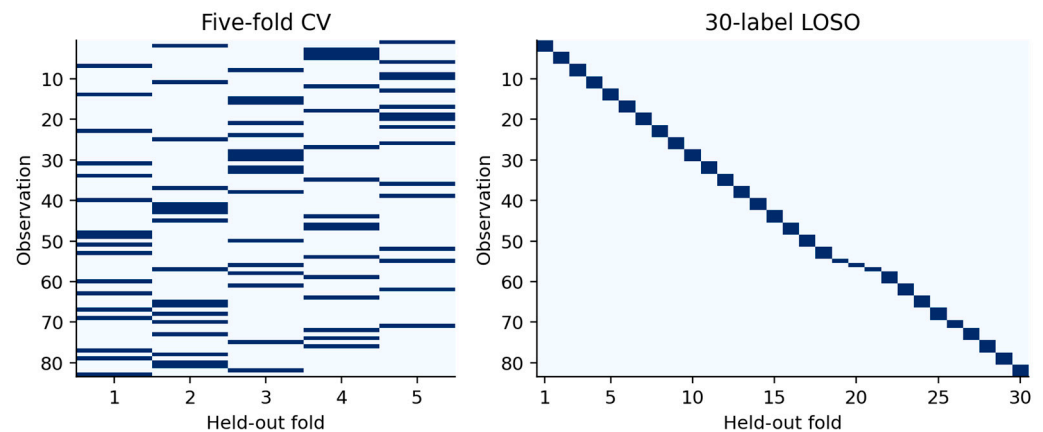


Figure 2. Shared held-out assignments for five-fold CV and 30-label LOSO. Dark cells identify test records. All 83 observations appear once as test records in each scheme.

2.5. Artificial Neural Network Baseline Model

Artificial neural networks (ANNs) are widely used for nonlinear regression problems due to their universal approximation capability and their ability to capture complex relationships between input variables and model outputs without requiring explicit assumptions about the underlying physical mechanisms. In the context of anaerobic digestion, neural network models have been successfully applied to predict methane yield, biogas production, and process performance based on substrate characteristics and operating conditions.

A single-hidden-layer feedforward ANN is used as the deterministic baseline. The principal model has 10 tanh hidden units and six linear outputs. A 20-unit version is evaluated in CV only to describe sensitivity to width; it does not replace the prespecified principal model.

The network maps the input vector defined in Equation (1) to the multi-output response vector defined in Equation (2). A single hidden layer feedforward architecture is adopted in order to balance model flexibility with numerical stability and interpretability. Networks with a single hidden layer are theoretically capable of approximating any continuous function given sufficient neurons, while remaining computationally efficient for relatively small datasets.

The mathematical form of the network mapping can be written as

$$\hat{y} = W_2 \phi(W_1 x + b_1) + b_2, \quad (14)$$

where W_1 and W_2 denote weight matrices connecting the input layer to the hidden layer and the hidden layer to the output layer, respectively, while b_1 and b_2 denote bias vectors. The function $\phi(\cdot)$ represents the nonlinear activation function applied to the hidden layer neurons, enabling the network to approximate nonlinear input–output relationships.

The ANN is trained for 400 full-batch Adam updates [30] with learning rate 0.01, moment coefficients 0.9 and 0.999, numerical epsilon 10^{-8} and L2 weight penalty 0.001. The objective is half the summed squared standardized-output residual divided by the number of training rows, plus half the L2 coefficient times the squared weight norm. Biases are not penalized.

Weights are initialized deterministically: first-layer entries are $0.25 \sin(1.7k)$ and second-layer entries are $0.25 \cos(1.3k)$, using consecutive one-based k in column-major order; biases start at zero. This initialization is reproduced directly in PYTHON (3.12.14 with NumPy 2.3.5, SciPy 1.17.0, pandas 2.2.3 i Matplotlib 3.10.8), MATLAB (24.1.0.3050227 (R2024a) Update 8). It is not an ensemble of random ANN restarts.

The model uses the designated external training set without an internal validation split, early stopping or test-driven hyperparameter tuning. Input and output normalization is fitted only on those training rows. The 400 updates are performed in every CV and LOSO iteration.

All principal ANN tables, figures and LOSO examples use the 10-unit Adam implementation. The 20-unit Adam model is a separate CV diagnostic, reported in Table A5. Neither configuration is presented as the result of an exhaustive architecture search or as a universally optimal ANN.

The ANN returns point predictions, without a probabilistic interval. Its 10- and 20-unit configurations are listed in Table 2. Differences between them are a diagnostic of this implementation and should not be generalized to all neural networks.

Table 2. ANN configurations in the analysis. The 10-neuron model is fixed for the principal CV and LOSO comparisons.

Training	Hidden Neurons	Activation	Role
Adam	10	tanh/linear	Principal model
Adam	20	tanh/linear	CV architecture diagnostic

2.6. Gaussian Process Regression with Uncertainty Quantification

GP regression supplies a predictive mean and a nominal probabilistic interval [24]. The predictive model uses a squared-exponential kernel on standardized continuous inputs with one fixed length scale of 2, unit signal variance and noise variance 0.05 in standardized-output units. A diagonal jitter of 10^{-10} is added for numerical factorization.

The kernel and noise parameters are prescribed assumptions, not estimates of experimental noise or optimized hyperparameters. Unit signal variance matches the scale of each standardized training target. Length scale 2 makes squared-exponential covariance decay to $\exp(-1/2)$ at standardized Euclidean distance 2. Noise variance 0.05 corresponds to a noise standard deviation of about 0.224 times the training-output standard deviation and regularizes the kernel system. These choices define a reproducible smooth baseline; they are not established as optimal. No marginal-likelihood optimization or parameter-sensitivity analysis is performed for this predictive GP. All six targets use the same input kernel with separate training-output standardization. The auxiliary ARD fit in Section 2.8.1 is distinct from the GP used for validation.

Formally, a Gaussian process is defined as

$$f(x) \sim GP(0, k(x, x')), \quad (15)$$

where $k(\cdot, \cdot)$ denotes the covariance or kernel function that defines the similarity between input points. The kernel function determines how strongly the predicted outputs at two input locations are correlated and therefore encodes assumptions about smoothness and variability of the underlying function.

The observation variance equals the conditional latent-function variance plus the specified observation-noise variance, transformed back using the training-output scale. The nominal 95% intervals are conditional on the fixed kernel, noise and fitted preprocessing assumptions. They do not propagate uncertainty in hyperparameter choice or estimation, and include no separate model-discrepancy or between-protocol uncertainty term. They therefore do not represent total predictive uncertainty.

In the present study, separate GP models are trained for each output variable. This approach avoids imposing explicit assumptions about correlations between methane and biogas yield measures while allowing uncertainty to be estimated independently for each

output. The resulting probabilistic predictions enable evaluation not only of prediction accuracy but also of the reliability of uncertainty estimates.

Interval reliability is evaluated using prediction interval coverage probability (PICP) and mean prediction interval width (MPIW). PICP is the fraction of held-out measured values inside the nominal interval:

$$\text{PICP}_\alpha = \frac{1}{N} \sum_{i=1}^N I(y_i \in [\mu_i - z_\alpha \sigma_i, \mu_i + z_\alpha \sigma_i]). \quad (16)$$

For central nominal coverage $\alpha = 0.95$, the multiplier z_α is 1.96. PICP is reported as a percentage in the tables.

Observed coverage close to the nominal level is desirable, but does not establish full calibration across outputs or input regions. In particular, a satisfactory pooled CV coverage may coexist with under-coverage for held-out substrate labels.

While PICP quantifies the coverage of uncertainty intervals, it does not describe their width. For this reason, the mean prediction interval width (MPIW) is also computed:

$$\text{MPIW}_\alpha = \frac{1}{N} \sum_{i=1}^N 2z_\alpha \sigma_i. \quad (17)$$

MPIW is reported in the units and mass basis of each output. Primary coverage summaries average the two fresh-mass output coverages; normalized-output coverage and width are reported separately in Table A6.

The PICP and MPIW metrics must be interpreted jointly. A model may achieve high coverage by producing excessively wide intervals, which would indicate poor predictive usefulness. Conversely, very narrow intervals may lead to low coverage, indicating over-confident predictions. A well-calibrated probabilistic model therefore aims to achieve high coverage while maintaining relatively narrow uncertainty intervals. Detailed per-output cross-validation results for Gaussian process regression, including PICP95 and MPIW uncertainty metrics, are provided in Table A6 (Appendix A).

An auxiliary ARD analysis uses dimension-specific lengths to describe the fitted response in the full dataset. Its role and limitations are distinguished from the fixed isotropic kernel in predictive validation.

2.7. Reduced-Order Multi-Output Modeling Using POD–RBF

POD–RBF represents the output matrix using a lower-dimensional linear basis. Strong correlations are expected because the six outputs share measured gas quantities and deterministic mass normalization. The method is evaluated as a compact representation, without assuming that compression guarantees better prediction.

To capture these dependencies, a reduced-order multi-output surrogate modeling framework is employed based on the combination of proper orthogonal decomposition (POD) and radial basis function (RBF) regression. This approach transforms the original multi-output regression problem into a lower-dimensional representation that captures the dominant patterns of variability in the output space.

For each training fold, let Y be its output matrix. The training output mean is subtracted without scaling the output columns, and SVD is applied to this centered training matrix:

$$Y - \bar{y} = U \Sigma V^\top, \quad (18)$$

where U , Σ , and V are matrices resulting from the decomposition. The columns of V represent orthogonal basis vectors (POD modes) spanning the output space, while the

diagonal elements of Σ correspond to singular values that quantify the importance of each mode.

Modes are ordered by singular value. The smallest rank retaining at least 99% of the summed squared singular values is selected using that training matrix alone. Table A7 reports held-out POD–RBF performance.

The amount of variance explained by the first r modes can be quantified using the cumulative variance ratio

$$\eta(r) = \frac{\sum_{i=1}^r \sigma_i^2}{\sum_{i=1}^M \sigma_i^2}. \quad (19)$$

Because the outputs are not scaled before POD, quantities with larger numerical variation carry greater weight in the energy criterion. The retained energy is a property of this chosen output representation, not an independent measure of biochemical dimensionality.

A Gaussian RBF regressor maps standardized inputs to the retained modal coefficients. It uses length scale 2 and diagonal ridge coefficient 0.01. Its coefficient matrix solves the regularized training kernel system. The input scaling, output mean, basis and rank are fitted using only the training data within each CV and LOSO fold.

After predicting the modal coefficients $\hat{a}$ for a new input point, the corresponding output vector is reconstructed using the retained POD modes. The reconstructed output can therefore be written as

$$\hat{y} = \hat{a}V_r^\top + \bar{y}, \quad (20)$$

where V_r denotes the matrix containing the retained POD modes.

POD reduces the number of response coordinates to be learned and constrains predictions to the retained training-output subspace. This is a form of structural regularization. It does not enforce exact methane-to-biogas ratios, positivity, or the nonlinear row-specific conversion among FM, DM and ODM yields.

The descriptive full-data modes shown later are fitted separately for visualization. They are never reused in held-out validation. The computational logs record the predictive rank and training membership of every POD–RBF fit.

2.8. Feature Relevance and Global Sensitivity Analysis

Feature diagnostics quantify the behavior of fitted surrogates. They do not identify causal effects in anaerobic digestion, especially where inputs are correlated or enter the arithmetic conversion of outputs. Three complementary analyses are retained:

- Kernel-based relevance derived from Gaussian process hyperparameters;
- Permutation-based importance measuring performance degradation after feature randomization;
- Variance-based global sensitivity indices computed using Sobol analysis.

2.8.1. ARD Relevance

The auxiliary ARD GP is fitted to standardized continuous inputs and each standardized output in the full dataset. It uses unit signal variance and fixed noise variance 0.05. Five length scales are fitted with a deterministic bounded coordinate search, separately for each output.

The squared-exponential covariance with dimension-specific lengths is

$$k(x, x') = \exp \left[-\frac{1}{2} \sum_{j=1}^5 \frac{(x_j - x'_j)^2}{\ell_j^2} \right] \quad (21)$$

where the signal variance is one and the five lengths correspond to PH, COND, TIME, DM and ODM. Each length starts at 2 and is constrained to [0.1, 10]. Log-space step factors

are 2, 1.5, 1.2 and 1.05, with up to eight sweeps per factor. Both directions are tested in predictor order and an improvement greater than 10^{-10} is accepted. The objective is the negative log marginal likelihood; this is a local bounded fit.

The length scale parameter plays a key role in determining how strongly the predicted output responds to variations in a given input dimension. Large values of l_d imply that the predicted function varies slowly along that dimension, indicating weak sensitivity to that input variable. Conversely, small values of l_d correspond to rapid variation of the function and therefore stronger dependence on that feature.

For interpretability, feature relevance is expressed using the inverse length scale

$$I_d \propto \frac{1}{l_d}, \quad (22)$$

Inverse lengths are normalized to sum to one within each output. A larger value indicates a shorter fitted response length in that standardized coordinate; it is not a direct physical importance measure.

2.8.2. Permutation Importance

While ARD-based relevance reflects the internal structure of the Gaussian process kernel, it does not directly quantify the effect of individual features on prediction accuracy. For this reason, an additional model-agnostic feature importance measure based on permutation testing is employed.

For each of five held-out CV folds, one continuous descriptor is permuted ten times within the test rows while the others remain fixed. The trained model is not refitted. The procedure breaks the descriptor associations with the target and with other predictors, so marginal permutations may create combinations outside the observed correlation structure.

Let M denote a trained surrogate model and let x_d represent the d -th input variable. The importance of that variable is quantified as the increase in prediction error after permutation,

$$\Delta \text{sMAPE}(d) = \text{sMAPE}(M; \text{perm}(x_d)) - \text{sMAPE}(M). \quad (23)$$

Both terms use the mean of six output-wise sMAPEs on the same held-out fold. Their difference is measured in percentage points, not as a percent change relative to the baseline error.

If the permuted feature has little influence on model predictions, the resulting increase in error will be small. Conversely, a substantial increase in prediction error indicates that the model strongly relies on that variable.

The same test-row permutation maps are used for ANN, GP and POD-RBF. Ten changes are averaged within each fold, followed by the mean and sample SD across the five fold means. These SDs are descriptive rather than confidence intervals, because the corresponding training sets overlap. SILA is not permuted because it is absent from all predictors.

2.8.3. Sobol Global Sensitivity

Sobol indices describe the variance of the fitted GP methane-FM response over a specified input distribution [31]. The GP is fitted to the full experimental dataset for this diagnostic; this fit is separate from validation. It is then evaluated at numerical integration points, which are not added to the experimental dataset.

The surrogate model can be represented as a function

$$Y = f(X_1, \dots, X_D). \quad (24)$$

The five inputs PH, COND, TIME, DM and ODM are sampled independently from continuous uniform distributions bounded by their observed minimum and maximum in Table 1. Exact unrounded bounds are exported. SILA is excluded. This product distribution is an explicit analysis assumption, not an assertion that the measured inputs are independent; it can include physically uncommon combinations.

The first-order Sobol index measures the fraction of output variance that can be attributed solely to variations in input X_d . It is defined as

$$S_d = \frac{\text{Var}(E[Y | X_d])}{\text{Var}(Y)}. \quad (25)$$

This index quantifies the direct contribution of a single input variable to the variability of the model output while ignoring interactions with other variables.

However, nonlinear systems often exhibit strong interactions between inputs. To capture these effects, the total Sobol index is introduced. This index measures the overall contribution of input X_d , including all interaction effects with other variables:

$$S_{T_d} = 1 - \frac{\text{Var}(E[Y | X_{\sim d}])}{\text{Var}(Y)}, \quad (26)$$

where $X_{\sim d}$ denotes the set of all input variables except X_d .

Under the independent-input decomposition, the difference between total-order and first-order effects represents interactions in the fitted response over the chosen product measure. It does not establish a biochemical mechanism in the measured population.

Two independent Monte Carlo base matrices are used at $N = 4096$ and $16,384$. First-order effects use the centered Saltelli estimator, and total-order effects use the squared-difference Jansen form described in [31]. For five inputs, each estimate requires $7N$ model evaluations. Two hundred paired-row bootstrap resamples produce 95% percentile intervals. Negative estimates are retained as finite-sample estimates rather than clipped. The two sample sizes provide a convergence check; the intervals quantify Monte Carlo error conditional on the fitted GP and input law, not all uncertainty in the study.

2.9. Computational Workflow and Reproducibility

The numerical analysis is executed in Python from the experimental workbook using the documented algorithms. Source row identities, formulas, partitions, predictions, fold parameters and diagnostic outputs are exported. The MATLAB implementation uses the same measured CSV, fold assignments, permutation maps and Monte Carlo integration designs.

Figure 3 summarizes the workflow. Records are assigned to training and test sets before any learned transformation. Every predictive model is then fitted to its training rows and used once for the corresponding held-out predictions. Full-data POD and ARD plots are explicitly separated from predictive validation.

CV and LOSO errors use the definitions and aggregation levels in Section 2.3. All manuscript tables and numerical figures are generated from the same prediction and diagnostic exports, preserving consistency between the results and their displayed summaries.

For each validation experiment, predictions are generated for all output variables and performance metrics are computed according to the definitions introduced in Section 2.3. In addition to deterministic accuracy metrics, probabilistic models also provide predictive uncertainty estimates that are evaluated using interval coverage and width measures.

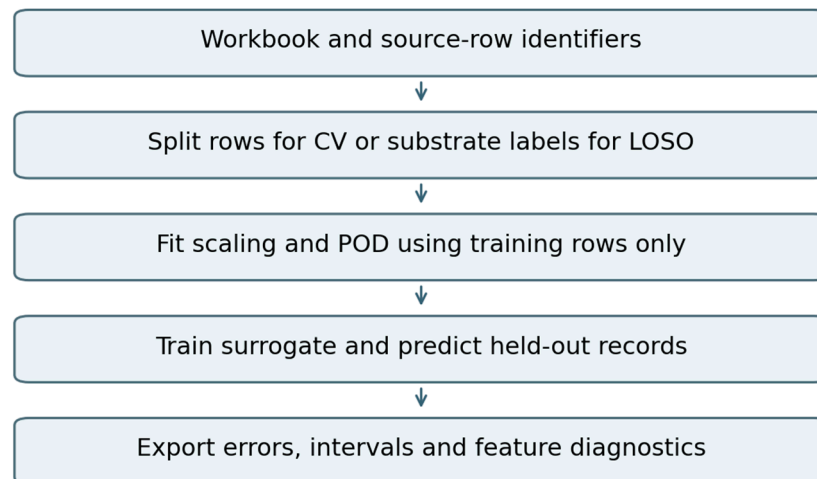


Figure 3. Computational workflow. Splitting precedes fitted preprocessing and POD. Auxiliary full-data diagnostics are separate from validation.

MATLAB source code is supplied for independent reproduction, but MATLAB was not executed in this study. Accordingly, all reported numerical results come from the executed Python workflow; cross-language agreement is not claimed on the basis of source-code availability alone.

The computational parameters and software versions are recorded in the run manifest. These choices make the comparison reproducible, while the narrow model configurations, single CV partition and limited experimental metadata remain limitations.

3. Results

The results derive from the 83 experimental records and the implementations specified in Section 2. They cover CV accuracy, uncertainty, reduced-order structure, LOSO performance and feature diagnostics.

Comparisons are descriptive. No statistical significance test between model families is performed, and smaller observed errors are not described as statistically significant.

3.1. Cross-Validation Performance

The first stage of model evaluation focuses on interpolation accuracy within the dataset distribution using the k-fold cross-validation procedure described in Section 2.4.1. In this setting, each model is repeatedly trained on subsets of the data and evaluated on unseen samples drawn from the same distribution.

Table 3 reports pooled held-out CV metrics for the two primary measured fresh-mass outputs. Table A5 gives the corresponding ANN width diagnostic. Tables A6 and A7 report each GP and POD-RBF output, including the four secondary normalized yields. The principal 10-unit ANN is used consistently in CV and LOSO.

Table 3. Primary comparison: five-fold CV for the two measured fresh-mass outputs. Metrics pool all 83 held-out predictions per output. RMSE is in $\text{m}^3/\text{Mg FM}$.

Model/Output	RMSE	R ²	MAPE (%)	sMAPE (%)
ANN/CH ₄ _fm	19.11	0.897	90.25	43.82
GP/CH ₄ _fm	16.96	0.919	45.49	32.21
POD-RBF/CH ₄ _fm	16.53	0.923	61.00	36.25
ANN/Biogas_fm	32.51	0.901	57.31	36.56
GP/Biogas_fm	31.59	0.906	37.54	27.21
POD-RBF/Biogas_fm	31.14	0.909	43.62	33.51

For the two primary outputs, mean CV RMSE is 25.81 m³/Mg FM for ANN, 24.27 for GP and 23.84 for POD-RBF. POD-RBF has the lowest RMSE for each fresh-mass output. GP has the lowest mean CV sMAPE (29.71%), followed by POD-RBF (34.88%) and ANN (40.19%). Thus, the preferred model depends on the selected metric. Figure 4 separates primary fresh-mass and secondary normalized-output errors.

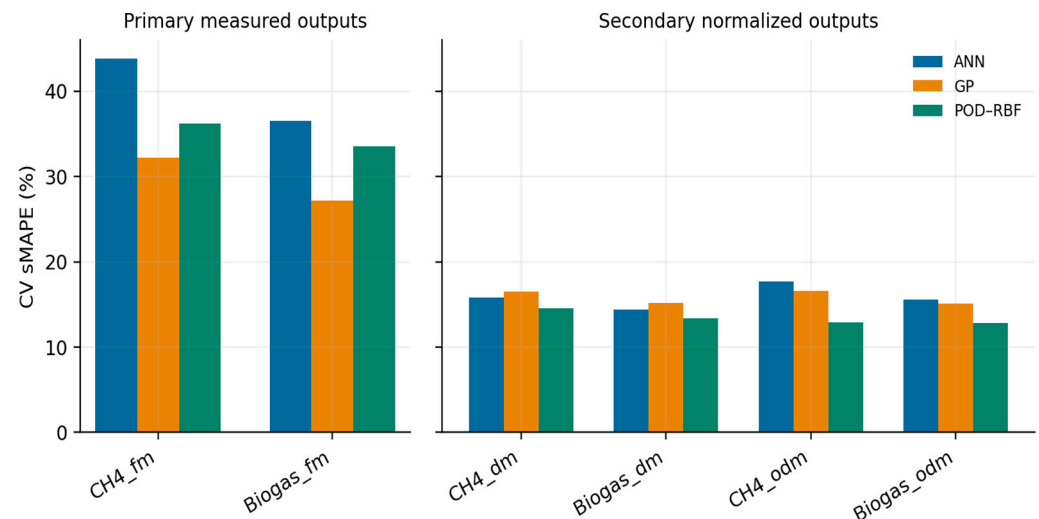


Figure 4. Per-output CV sMAPE for ANN, GP and POD-RBF from pooled held-out predictions. The two measured fresh-mass endpoints are primary; the four normalized endpoints are secondary.

Absolute and relative errors emphasize different parts of the output distribution. RMSE weights large deviations strongly, whereas percentage errors emphasize proportional discrepancies. The primary RMSE comparison uses the common fresh-mass basis; secondary normalized-output errors are reported separately. All rankings remain conditional on the single CV partition.

Figure 5 shows all five methane-FM fold sMAPEs for ANN, GP and POD-RBF. The scatter points display the small number of fold observations behind the descriptive boxplots.

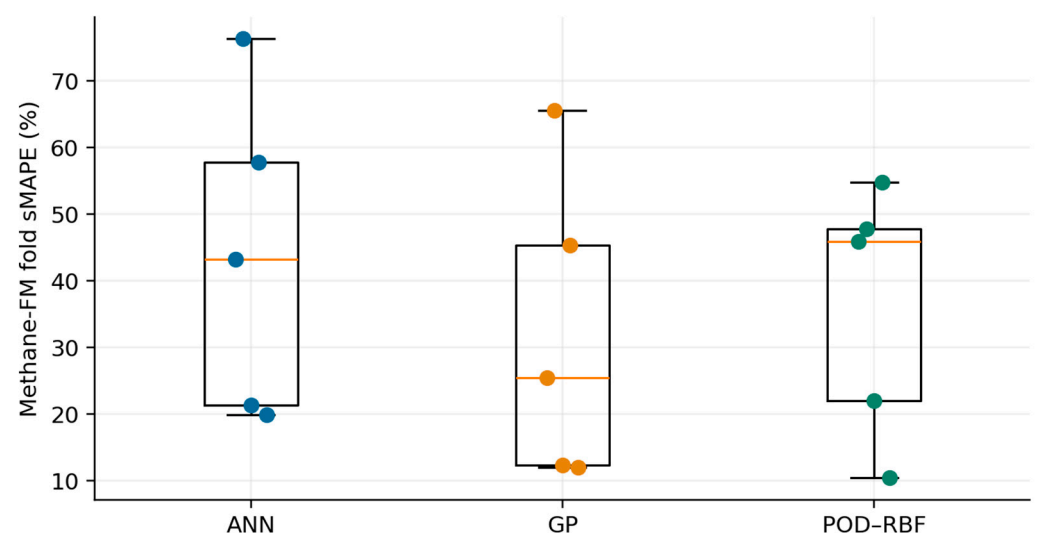


Figure 5. Methane-FM CV fold sMAPE for all three principal models. Points show the five fold values; boxplots are descriptive and do not represent independent experiments.

The fold errors vary for each implementation. Their distribution does not establish that one model family is intrinsically more stable, and five overlapping training sets do

not provide independent experimental replicates. Complete group-wise LOSO errors are reported in Table A4.

3.2. Gaussian Process Predictions and Uncertainty Estimation

A key advantage of Gaussian process regression is its ability to provide predictive uncertainty estimates in addition to point predictions. This property allows model confidence to be evaluated alongside predictive accuracy.

Figure 6 shows held-out GP methane-FM predictions with individual nominal 95% observation intervals, sorted by the measured yield for display. Unlike a band based on average width, these intervals use the predictive standard deviation for each held-out record.

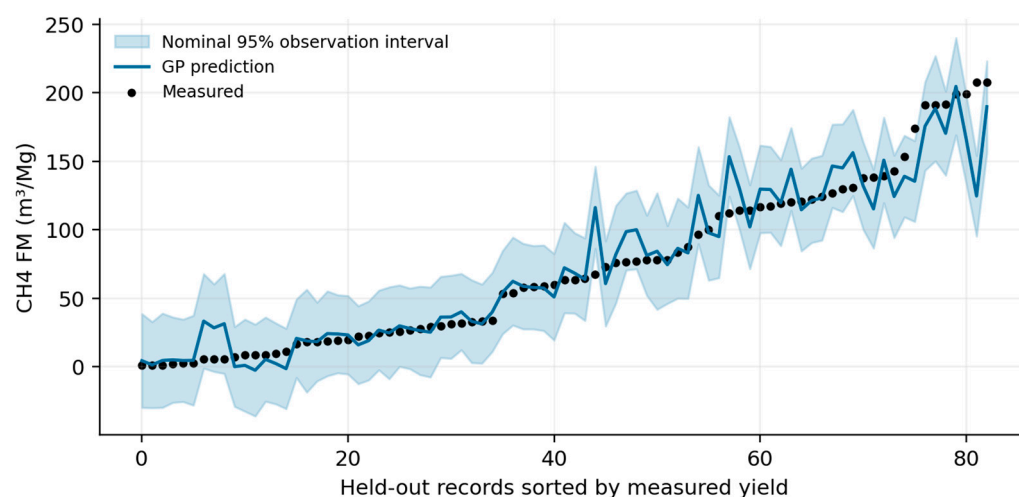


Figure 6. Held-out GP methane-FM predictions with individual conditional nominal 95% observation intervals. Intervals include prescribed observation noise but omit hyperparameter and model-discrepancy uncertainty. Records are sorted by measured yield only for display.

The interval widths follow the fixed kernel and observation-noise assumptions. A wide or narrow interval alone does not demonstrate correct uncertainty quantification. Calibration must be assessed against held-out measurements and may change between CV and label holdout.

For the primary endpoints, CV PICP95 is 93.98% for methane-FM and 92.77% for biogas-FM, averaging 93.37%. LOSO coverage is 75.90% and 78.31%, respectively, averaging 77.11%. Secondary normalized-output CV coverage ranges from 74.70% to 84.34% (Table A6), showing that near-nominal fresh-mass coverage does not extend to every reported output. The intervals are not established as fully calibrated.

3.3. Reduced-Order Structure of the Multi-Output Space

The multi-output nature of the prediction problem allows correlations between methane and biogas yields to be exploited using dimensionality reduction techniques. The proper orthogonal decomposition described in Section 2.7 provides insight into the structure of the output space.

Figure 7 presents the first three descriptive POD modes of the full centered, unscaled output matrix. Their signs are fixed by making the largest-magnitude coefficient positive for reproducible plotting; mode signs have no physical interpretation.

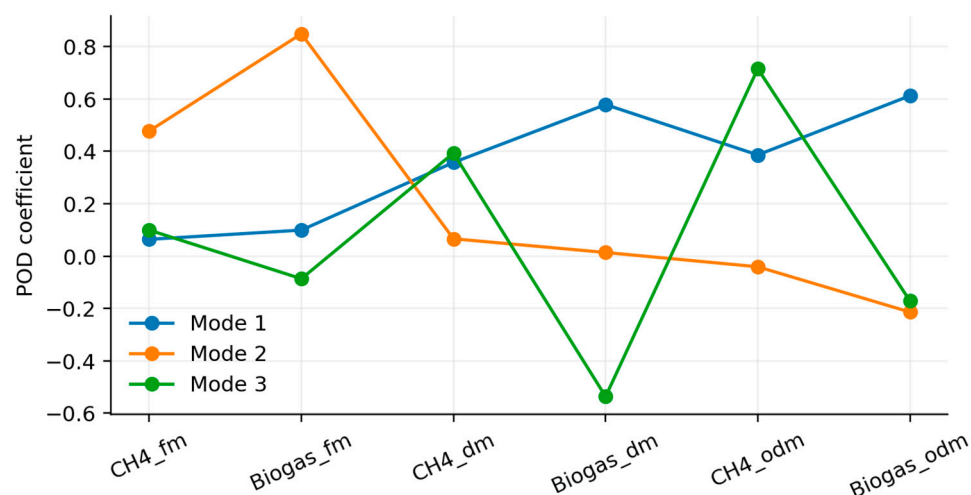


Figure 7. First three descriptive POD modes of the full centered, unscaled output matrix. These full-data modes are not reused in predictive validation.

The corresponding cumulative variance explained by successive modes is shown in Figure 8 and in Table 4.

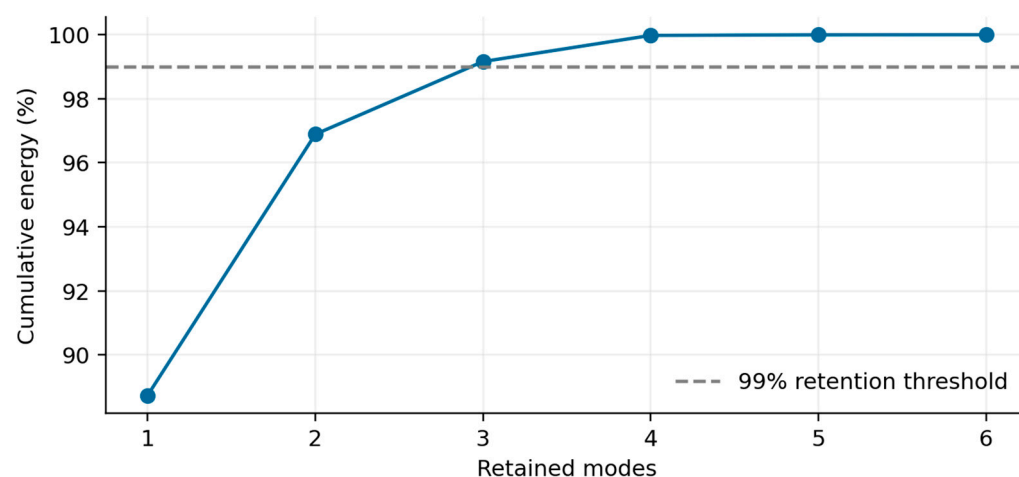


Figure 8. Cumulative POD energy for the full output matrix. The dashed line marks 99%; each validation fold independently selects its own training basis and rank.

Table 4. Descriptive POD energy of the full centered, unscaled output matrix. Predictive POD bases are fitted separately on training rows.

Modes	Cumulative Energy
1	0.887358
2	0.968908
3	0.991621
4	0.999785
5	0.999969
6	1.000000

The first three full-data modes retain 99.16% of centered output energy; two modes retain 96.89%. Under the 99% training-only criterion, 33 of the 35 predictive POD fits retain three modes and two retain four. The compactness partly reflects mass-conversion identities and the chosen unscaled output representation. It does not by itself establish an independent multi-output advantage.

3.4. Generalization Performance: LOSO Validation

LOSO predicts all observations of one excluded literal feedstock label using the remaining labels. The comparison is more demanding than random-row CV, while retaining the limitations of label identity and small group sizes described in Section 2.4.2.

Figure 9 displays the two-output fresh-mass group sMAPEs, on a common 0–200% scale for all models. Three singleton labels and one two-record label remain included. Secondary six-output summaries are reported in Tables A4 and A8.

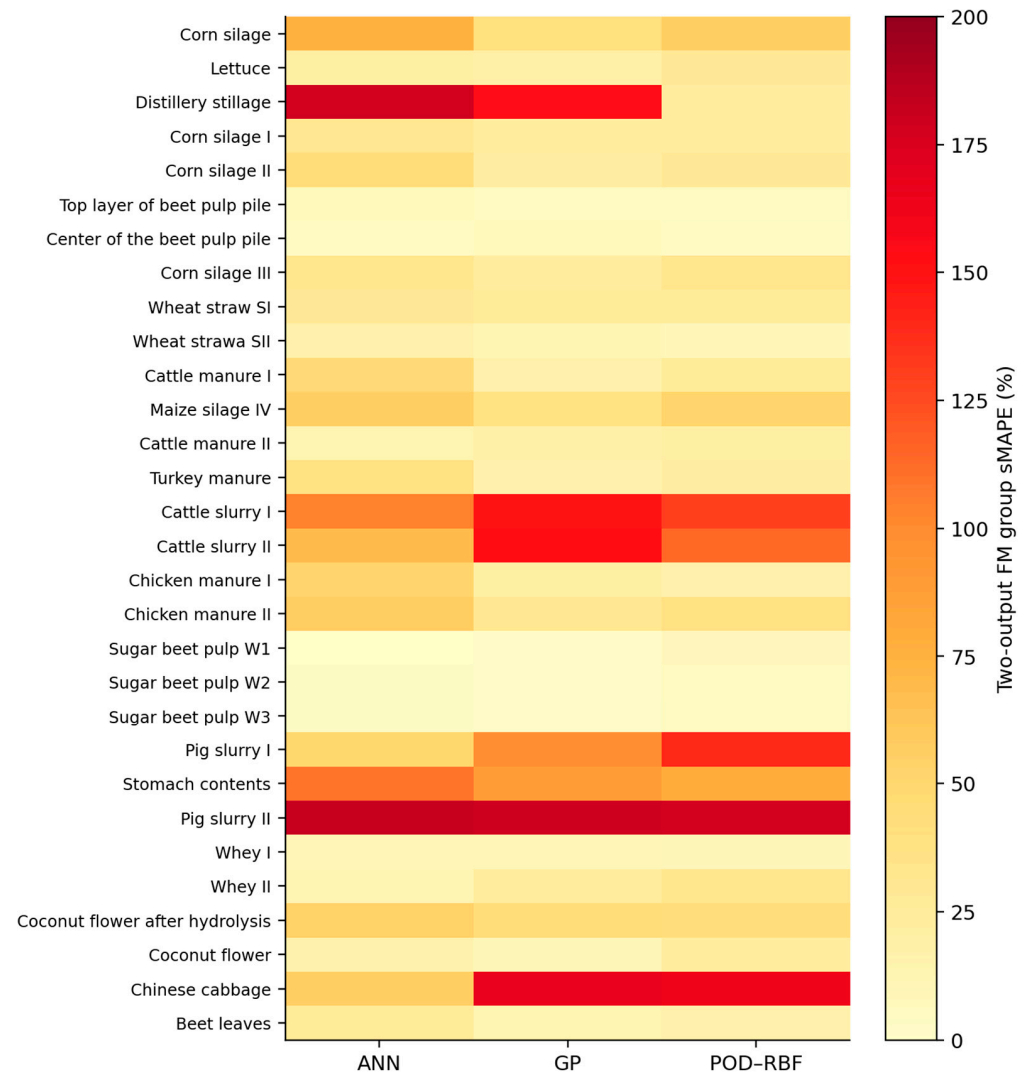


Figure 9. Primary LOSO group sMAPE, averaged over methane-FM and biogas-FM, for all 30 supplied labels.

Figure 10 sorts group errors separately within each model. Thus, the same horizontal rank may refer to different substrate labels; identities should be compared in Figure 9.

Figure 11 identifies the lowest group sMAPE among the three models for each label. These counts describe the supplied dataset and do not constitute a statistical comparison.

Prediction errors vary strongly across labels. For Pig slurry II, primary group sMAPE is 180.99% for ANN, 178.37% for GP and 176.99% for POD-RBF. Distillery stillage gives 177.68% for ANN and 152.49% for GP, compared with 23.05% for POD-RBF. These large bounded percentage errors reveal substantial prediction failures and preclude a claim of robust extrapolation.

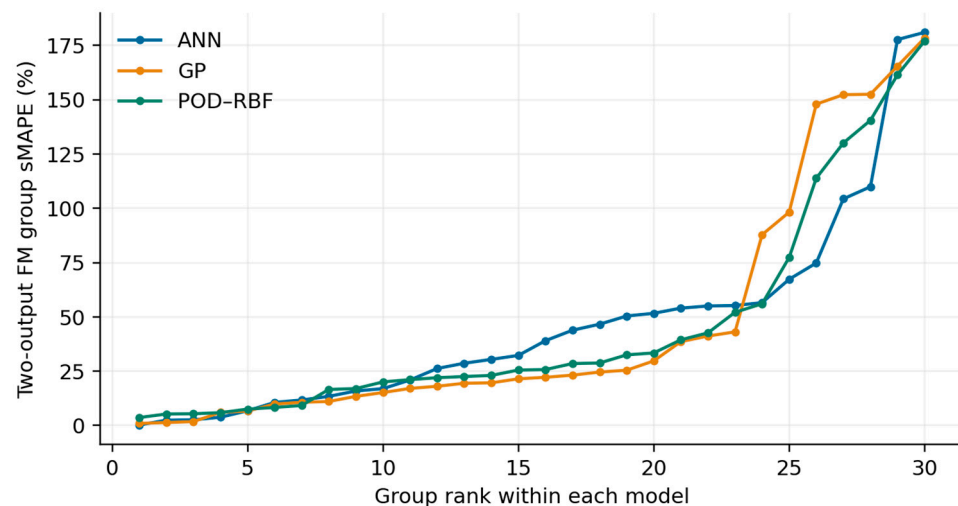


Figure 10. Primary two-output fresh-mass LOSO group errors sorted separately within each model. Equal ranks need not identify the same feedstock label.

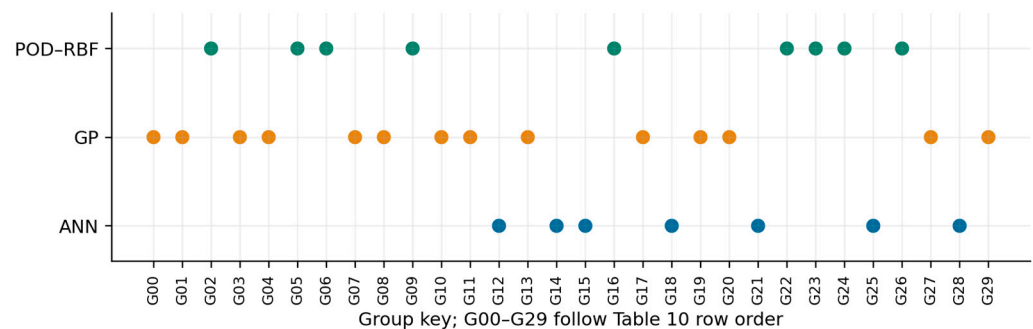


Figure 11. Lowest-error model by primary two-output fresh-mass group sMAPE; the counts are descriptive.

Table 5 summarizes the 30 group sMAPEs with equal label weights. The mean, median and sample SD describe different aspects of the distribution and should not be interchanged with statistics pooled over all 83 rows.

Table 5. LOSO sMAPE across 30 literal substrate labels. Each group error averages the two fresh-mass output-wise errors; every group receives equal weight.

Model	Mean sMAPE (%)	Median sMAPE (%)	Group SD (%)
ANN	46.35	35.65	45.84
GP	46.78	21.84	55.75
POD-RBF	45.09	25.64	48.99

All 83 observations contribute exactly once to held-out predictions. The 30 groups contain 26 sets of three, one set of two and three singletons. No small group is removed from the mean or winner count.

Primary mean group sMAPE is 46.35% for ANN, 46.78% for GP and 45.09% for POD-RBF. POD-RBF has the lowest mean, but the differences are descriptive and are not statistically established advantages. GP has the lowest median group error (21.84%), compared with 25.64% for POD-RBF and 35.65% for ANN, so ordering also depends on the summary statistic.

ANN has the lowest primary group error for seven labels, GP for 14 and POD-RBF for nine (Figure 11). The secondary six-output winner counts differ (Table A8). No

model wins every group, and the reduced basis does not guarantee uniformly superior holdout performance.

Table 6 distinguishes the output representations and uncertainty capabilities of the implementations. Neither a shared ANN layer nor a POD subspace is equivalent to exact enforcement of mass-normalization identities.

Table 6. Structural comparison of the implemented models. Shared outputs or a reduced basis do not enforce the row-specific mass conversions exactly.

Model	Output Representation	Probabilistic Interval	Mass Identities
ANN	Shared hidden layer	None	Not enforced
GP	Six independent outputs	Nominal Gaussian	Not enforced
POD–RBF	Training-output subspace	None	Not enforced exactly

3.5. Feature Relevance Based on Gaussian Process ARD

Figure 12 reports normalized inverse lengths from the auxiliary full-data ARD analysis. This diagnostic is separate from the fixed-kernel GP evaluated in CV and LOSO.

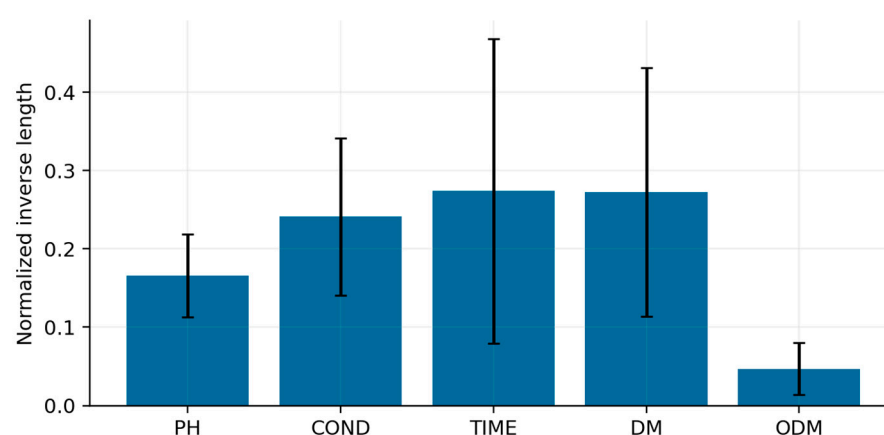


Figure 12. Auxiliary ARD relevance. Bars show mean normalized inverse lengths over six outputs; error bars show sample SD across outputs, not confidence intervals.

TIME has the largest mean auxiliary ARD score (0.2739), closely followed by DM (0.2724), and then COND (0.2410). The sample SD across outputs shows that fitted coordinate lengths differ among yield bases. The bounded local search and input correlations limit physical interpretation (Table 7).

Table 7. Auxiliary full-data ARD diagnostics. Inverse lengths are normalized within each output; mean and sample SD are across six outputs, not uncertainty intervals.

Descriptor	Mean Relevance	SD Across Outputs
PH	0.1657	0.0531
COND	0.2410	0.1000
TIME	0.2739	0.1944
DM	0.2724	0.1586
ODM	0.0470	0.0331

3.6. Permutation-Based Feature Importance

Figures 13–15 report held-out permutation importance for GP, POD–RBF and ANN, respectively. Each bar is the mean of five fold means, and the error bar is their sample SD. Every fold mean is based on ten saved test-row permutations.

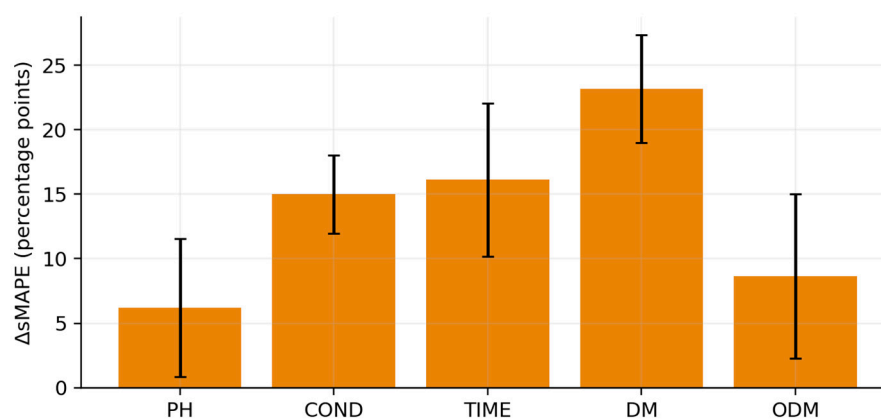


Figure 13. GP held-out permutation importance in percentage points of sMAPE. Bars show five fold means averaged over ten permutations each; error bars show SD across fold means.

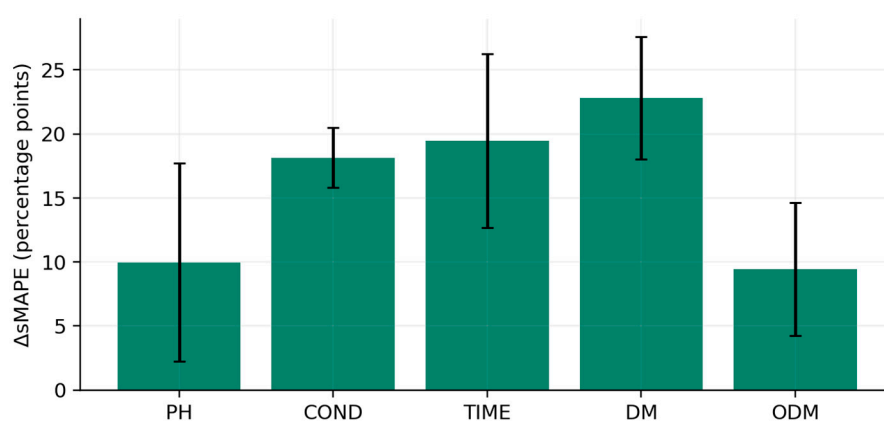


Figure 14. POD-RBF held-out permutation importance with the same definition and permutations as in Figure 13.

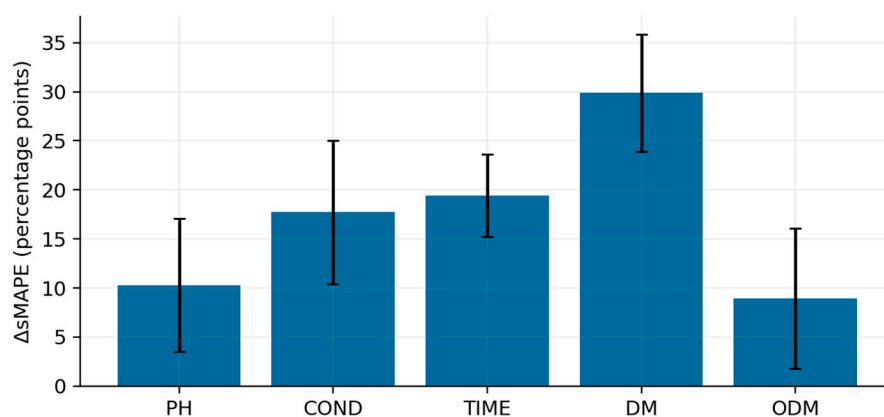


Figure 15. ANN held-out permutation importance with the same definition and permutations as in Figure 13. The principal 10-neuron model is used.

DM produces the largest mean permutation-induced error increase for all three implementations, followed by TIME. These secondary six-output diagnostics combine measured and normalized targets. DM is explicitly involved in converting outputs between mass bases, and TIME varies with the recorded experiments; their rankings do not identify isolated biochemical effects.

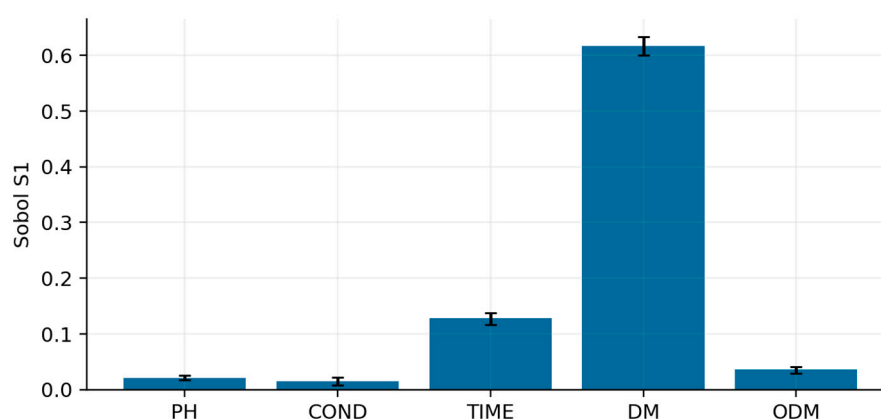
Table 8 provides the numerical comparison for all three models in percentage points of sMAPE. Because marginal permutations may violate input correlations, these values measure model dependence under a specific perturbation rather than isolated causal effects.

Table 8. Held-out permutation importance. Values are mean $\pm$ sample SD of five fold means, each based on ten permutations. Units are percentage points of six-output sMAPE.

Descriptor	ANN Δ sMAPE	GP Δ sMAPE	POD-RBF Δ sMAPE
PH	10.29 $\pm$ 6.78	6.21 $\pm$ 5.35	9.97 $\pm$ 7.74
COND	17.72 $\pm$ 7.30	15.01 $\pm$ 3.03	18.15 $\pm$ 2.32
TIME	19.42 $\pm$ 4.18	16.13 $\pm$ 5.93	19.47 $\pm$ 6.78
DM	29.89 $\pm$ 5.99	23.19 $\pm$ 4.19	22.81 $\pm$ 4.79
ODM	8.92 $\pm$ 7.14	8.64 $\pm$ 6.39	9.44 $\pm$ 5.18

3.7. Global Sensitivity Analysis

Global sensitivity describes the fitted GP methane-FM response over independent uniform variations of the five descriptors. It does not add observations to the experimental dataset. Figures 16 and 17 use $N = 16,384$, with paired-row bootstrap intervals.

**Figure 16.** First-order GP methane-FM Sobol indices at $N = 16,384$ under independent uniform descriptor sampling. Error bars show 95% paired-row bootstrap intervals.**Table 9.** GP methane-FM Sobol estimates under independent uniform descriptor sampling. Brackets give paired-row bootstrap 95% intervals using 200 resamples.

N	Input	S1 [95% Interval]	ST [95% Interval]
4096	PH	0.0173 [0.0082, 0.0242]	0.0509 [0.0478, 0.0543]
4096	COND	0.0005 [−0.0156, 0.0129]	0.1059 [0.0987, 0.1141]
4096	TIME	0.1102 [0.0880, 0.1299]	0.2596 [0.2436, 0.2750]
4096	DM	0.6447 [0.6092, 0.6795]	0.7238 [0.6873, 0.7568]
4096	ODM	0.0337 [0.0215, 0.0466]	0.0859 [0.0800, 0.0909]
16,384	PH	0.0213 [0.0174, 0.0252]	0.0495 [0.0482, 0.0512]
16,384	COND	0.0154 [0.0083, 0.0220]	0.1008 [0.0978, 0.1043]
16,384	TIME	0.1284 [0.1169, 0.1376]	0.2604 [0.2538, 0.2678]
16,384	DM	0.6173 [0.6008, 0.6337]	0.7056 [0.6876, 0.7240]
16,384	ODM	0.0357 [0.0294, 0.0412]	0.0898 [0.0869, 0.0923]

Figure 16 presents the first-order Sobol indices (S1), which represent the direct contribution of each input variable to the variance of the predicted outputs. In contrast, Figure 17 shows the total Sobol indices (ST), which account for both the direct effects of individual variables and their interaction effects with other inputs.

At $N = 16,384$, DM has first-order index 0.6173 and total-order index 0.7056. TIME follows with 0.1284 and 0.2604. These values are conditional on the chosen GP and rectangular input domain; the difference between total and first order reflects interactions within that numerical model.

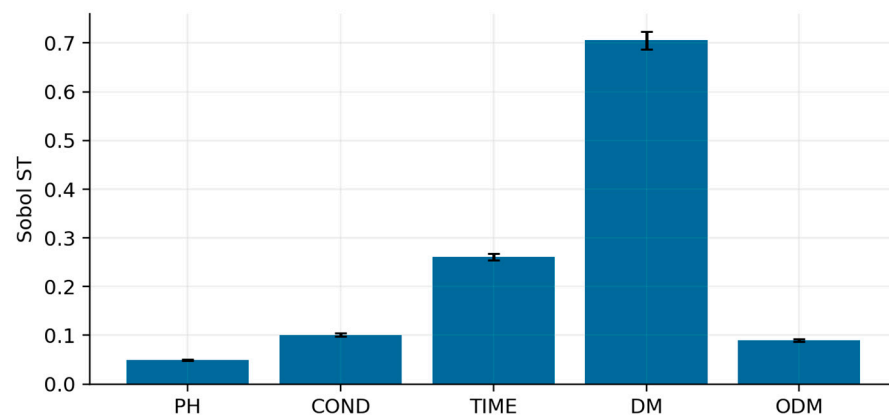


Figure 17. Total-order Sobol indices for the same model, sample design and input law as in Figure 16. Labels and values match Table 9.

Table 9 includes both sample sizes. The COND first-order estimate is 0.0005 at $N = 4096$, with a 95% bootstrap interval of $[-0.0156, 0.0129]$, and 0.0154 at $N = 16,384$. All reported first-order and total-order point estimates are non-negative. Negative bootstrap endpoints are retained as finite-sample estimation effects, not interpreted as negative variance contributions. The intervals and sample-size comparison describe conditional Monte Carlo variability, not experimental causality or guaranteed complete convergence.

3.8. Summary of Surrogate Model Performance

The comparison produces no universal winner. For the primary fresh-mass endpoints, POD-RBF has the lowest mean CV RMSE and mean group LOSO sMAPE, whereas GP has the lowest CV sMAPE and median group LOSO sMAPE. These descriptive rankings depend on the error criterion and validation setting; output compression alone does not explain or establish a general predictive advantage.

Table 10 lists all labels, sample counts and group sMAPEs. The error increase from CV to LOSO and the extreme errors for some labels restrict claims of reliable prediction beyond the available feedstock conditions.

Table 10. Primary LOSO comparison for all 30 supplied labels and 83 held-out records. Each value averages methane-FM and biogas-FM sMAPE.

Substrate Label	n	ANN sMAPE (%)	GP sMAPE (%)	POD-RBF sMAPE (%)
Corn silage	3	74.80	41.17	56.03
Lettuce	3	20.94	19.45	28.77
Distillery stillage	3	177.68	152.49	23.05
Corn silage I	3	30.43	22.18	22.54
Corn silage II	3	43.87	21.50	28.53
Top layer of beet pulp pile	3	6.79	6.01	5.29
Center of the beet pulp pile	3	3.80	6.65	3.66
Corn silage III	3	32.27	23.18	33.40
Wheat straw SI	3	28.66	25.45	25.53
Wheat straw SII	3	16.96	11.07	8.30
Cattle manure I	3	46.63	17.08	25.74
Maize silage IV	3	55.01	38.61	52.11
Cattle manure II	3	13.42	18.04	20.06
Turkey manure	3	39.04	15.17	21.11
Cattle slurry I	3	104.32	147.93	130.05
Cattle slurry II	3	67.32	152.29	113.84

Table 10. Cont.

Substrate Label	n	ANN sMAPE (%)	GP sMAPE (%)	POD-RBF sMAPE (%)
Chicken manure I	3	51.65	19.64	17.00
Chicken manure II	3	56.56	29.74	39.44
Sugar beet pulp W1	1	0.24	0.96	7.54
Sugar beet pulp W2	1	2.45	1.80	5.91
Sugar beet pulp W3	1	2.61	1.35	5.40
Pig slurry I	3	50.41	98.22	140.48
Stomach contents	3	109.91	87.84	77.37
Pig slurry II	3	180.99	178.37	176.99
Whey I	3	10.63	10.67	9.24
Whey II	2	11.73	24.60	32.52
Coconut flower after hydrolysis	3	54.03	43.10	42.65
Coconut flower	3	15.90	9.87	22.00
Chinese cabbage	3	55.23	165.46	161.43
Beet leaves	3	26.23	13.42	16.59

GP intervals and feature diagnostics supply additional information, but coverage deteriorates under label holdout and the sensitivity results depend on representation and sampling assumptions. These limitations guide the interpretation below.

4. Discussion

The comparison addresses a setting previously studied using machine learning for biogas prediction [11–13,17,18]. Its contribution is the common evaluation of three specified surrogate representations, with explicit group holdout, coverage and output-dependence checks. Differences from published model performance cannot be interpreted as superiority because datasets and validation designs differ.

Table 11 summarizes the numerical comparison. It separates predictive error from nominal uncertainty and output coupling. Industrial suitability cannot be established from this dataset.

Table 11. Summary of primary performance for the two fresh-mass outputs. CV entries average their two output-wise metrics; LOSO averages group means. RMSE is in $\text{m}^3/\text{Mg FM}$. The rankings depend on the error measure and validation protocol.

Criterion	ANN	GP	POD-RBF
Mean CV RMSE (FM)	25.81	24.27	23.84
Mean CV sMAPE (%)	40.19	29.71	34.88
Mean group LOSO sMAPE (%)	46.35	46.78	45.09
GP coverage CV/LOSO	–	93.37%/77.11%	–
Output coupling	Shared hidden layer	Independent outputs	Training POD basis
External industrial validation	Not performed	Not performed	Not performed

POD-RBF does not provide the lowest errors across all metrics or validation settings. The ranking changes with the error measure and prediction task.

4.1. Predictive Accuracy Under Interpolation Conditions

For the primary endpoints, POD-RBF gives mean CV RMSE $23.84 \text{ m}^3/\text{Mg FM}$ and mean R^2 0.916. GP gives the lowest mean relative errors among the principal models (MAPE 41.52% and sMAPE 29.71%). ANN has mean RMSE 25.81 and sMAPE 40.19%. These results concern one fixed partition; another partition could change the ordering. POD has useful approximation properties in other engineering applications [25,26], but the present comparison does not isolate output coupling from deterministic normalization.

The ANN20 diagnostic lowers primary CV sMAPE from 40.19% to 32.99% and RMSE from 25.81 to 24.30 m³/Mg FM (Table A5). This indicates width dependence in this implementation. The principal 10-unit model remains fixed for both validation schemes; the 20-unit diagnostic is not presented as an independently selected or validated optimal architecture.

The predictive GP is a fixed, isotropic kernel model. Its results characterize that implementation, rather than a comprehensively optimized GP family. Conversely, the ANN is one deterministic initialization and training schedule. Broader tuning or repeated initialization would address a different comparison and is not claimed in this study.

Random-row CV can benefit from having records with the same or related feedstock labels in training and test sets. Consequently, these results provide an interpolation-like assessment within the compiled dataset rather than a reliable estimate for novel feedstocks.

4.2. Generalization to Unseen Substrates

LOSO removes one entire supplied label at a time and evaluates predictions from continuous descriptors. The complete results cover all 83 records. Tables A1–A3 illustrate three distinct measured records and their held-out ANN predictions.

Mean primary group LOSO sMAPE of approximately 45–47% and much larger errors for some labels indicate substantial transfer limitations. The large percentage errors cannot be dismissed solely as a property of small denominators: bounded sMAPE also exposes pronounced prediction failures. No prospective or independent industrial dataset has been evaluated.

POD–RBF has the lowest primary mean group LOSO sMAPE, but its advantage is small and not statistically established. Its basis summarizes training-output variation and cannot supply omitted composition or experimental-condition information. It also does not enforce positivity or exact mass conversions.

GP reverts toward its prior behavior away from represented regions, but nominal predictive variance does not guarantee adequate coverage under a distribution change. The observed LOSO coverage shortfall is therefore considered together with point error, not used as reassurance about extrapolation.

The primary mean LOSO sMAPEs differ by less than two percentage points across the three implementations and do not justify a statistical ranking. Median errors and winning labels differ, and each model fails substantially for some feedstocks. Validation design and the intended error criterion are therefore central to interpretation.

Literal SILA labels are an operational grouping rule. Related maize/corn labels and W1–W3 records may share material or campaign characteristics; without independent batch metadata, LOSO cannot prove transfer to wholly independent experimental sources.

4.3. Role of Uncertainty Quantification in Surrogate Modeling

Nominal GP intervals are useful model outputs only if their coverage and width are evaluated under the intended prediction setting. Figure 6 and Table A6 provide individual intervals and per-output CV coverage; pooled LOSO coverage is also reported.

For the primary fresh-mass endpoints, mean coverage falls from 93.37% in CV to 77.11% in LOSO. Secondary normalized-output CV coverages are lower still than the primary CV values (see Table A6). Thus, near-nominal average fresh-mass coverage does not establish full calibration across outputs or held-out labels. Intervals include the fixed observation-noise term but omit hyperparameter uncertainty and any explicit model-discrepancy or between-protocol term.

ANN and POD–RBF provide no probabilistic uncertainty intervals in this implementation. The reduced POD subspace is a structural constraint, not an uncertainty estimate or a guarantee of physical consistency.

Any future use in operational decisions would require evaluation of interval behavior with independent data and fuller measurement metadata. The current intervals are conditional on the specified model and supplied descriptors.

4.4. Interpretation of Feature Relevance and Sensitivity

ARD, permutation importance and Sobol indices answer different questions. ARD summarizes fitted coordinate lengths, permutation importance measures held-out error changes, and Sobol analysis allocates model-output variance over a specified product distribution. Their rankings need not coincide.

The Sobol analysis targets methane yield on a fresh-mass basis, so the DM effect describes a relationship learned by this GP; it is not directly imposed by dividing that target by DM. In contrast, the multi-output POD, ARD and permutation diagnostics also include dry-mass targets that explicitly depend on DM and ODM. Neither type of diagnostic isolates a biochemical causal effect.

Correlated predictors also complicate the permutation analysis. Shuffling one column while holding the others fixed may create unusual combinations, so the resulting error increase describes dependence of this predictor on the observed associations and perturbation scheme.

The Monte Carlo intervals quantify sampling variability conditional on the fitted GP. They do not include uncertainty in the experimental measurements, the selected model family, the uniform input law or the bounds. Comparison of two sample sizes addresses numerical stability only within those assumptions.

SILA is not included in any predictor or feature diagnostic. This removes arbitrary categorical distances and avoids attributing physical meaning to derivatives or Sobol effects of integer labels.

4.5. Implications for Surrogate Modeling of Anaerobic Digestion Systems

The practical result is a transparent comparison whose claims are limited by measured-data coverage and model assumptions. It supports examining alternative surrogates and validation designs, but does not establish a deployable predictor for industrial digestion.

POD–RBF is useful when a compact representation is required, provided that all basis construction remains inside the training folds. In this application, deterministic output conversions make energy concentration a limited basis for claiming an intrinsic multi-output advantage.

GP supplies nominal probabilistic information, while ANN and POD–RBF supply deterministic predictions. These differences should be evaluated against the use case and held-out errors rather than converted into general quality labels.

Real-time optimization, digital twins and industrial decision support remain future possibilities. They would require independent and prospective validation, relevant on-line descriptors, and assessment of physical feasibility and uncertainty under changing operating conditions.

4.6. Limitations of the Study

The dataset comprises 83 observations with only one to three records per literal label. This limits the precision of group-wise errors and representation of the target population. Only one fixed five-fold partition is evaluated; metrics and apparent model rankings may depend on that split. Repeated-CV variability is unknown, and five fold scores do

not estimate uncertainty across alternative partitions. No repeated ANN initialization or statistical significance test is reported.

The source workbook and the data contributors establish the origin, recorded variables and mass conversions. However, detailed experiment-level metadata and a source-to-laboratory mapping are not retained in the compiled table. Protocol comparability is therefore a contributor-reported feature rather than a verified equivalence of every laboratory condition.

Only five continuous descriptors are available. Composition variables such as carbon-to-nitrogen ratio, lignocellulosic fractions and particle size are not included. TIME is the recorded fermentation duration and must be known for the stated prediction task; performance does not establish early prediction when this duration is unavailable.

The prescribed settings are not exhaustive optima. The predictive GP parameters are neither optimized nor subjected to sensitivity analysis, and its intervals condition on these fixed assumptions. GP treats outputs independently; POD weights unscaled output variation; ANN and POD-RBF can predict negative values. Four targets are deterministic transformations, so primary performance is based on the two measured fresh-mass endpoints, with normalized-output results treated as secondary.

LOSO uses the literal labels rather than independently documented experimental campaigns. Related labels and singleton W-labeled records limit the interpretation of unseen-feedstock performance. Sobol analysis uses independent uniform sampling despite possible dependence in the measurements, and its intervals reflect only conditional Monte Carlo error.

No prospective experiment or external industrial validation is reported. Further work could address richer descriptors, independent campaigns, physically constrained predictions and uncertainty calibration. These needs define the scope of the present conclusions rather than completed validation activities.

5. Conclusions

This study compares three surrogate implementations on 83 author-supplied experimental records using a common five-descriptor input space, five-fold CV and complete 30-label LOSO validation. Source identities, prediction exports and algorithms are available for reproduction.

First, primary performance is metric dependent. POD-RBF has the lowest mean CV RMSE (23.84 m³/Mg FM), whereas GP has the lowest mean CV sMAPE (29.71%). The single fixed CV partition leaves split dependence unquantified. A separate 20-unit ANN diagnostic shows sensitivity to network width.

Second, mean primary group LOSO sMAPE is 46.35% for ANN, 46.78% for GP and 45.09% for POD-RBF. These differences are descriptive; severe errors for several labels do not support generally robust prediction for unseen feedstocks.

Third, conditional nominal GP 95% observation coverage averages 93.37% in CV and 77.11% in LOSO for the primary endpoints. Lower secondary-output coverage and the LOSO shortfall show that the intervals are not established as fully calibrated. Hyperparameter and model-discrepancy uncertainties are not propagated.

Fourth, three full-data POD modes retain 99.16% of centered output energy, with three or four modes retained in predictive folds. Compactness partly reflects deterministic mass conversions and does not by itself establish a multi-output advantage or exact physical consistency.

Fifth, descriptor relevance depends on the model and analysis definition. The Sobol results identify DM and TIME as influential under the specified independent uniform input law; they do not establish biochemical causality.

Finally, excluding SILA from prediction removes arbitrary category distances while retaining it as a transparent grouping variable. The data and code support reproducibility of the numerical comparison, and the documented protocol and validation limitations restrict its current use to the stated research setting.

Future work should use independently documented experimental campaigns and prospective or external data to assess practical prediction reliability before considering industrial optimization or digital-twin applications.

Author Contributions: Conceptualization, K.W., N.K. and V.R.; methodology, K.W. and T.G.; software, T.G.; validation, N.K., V.R. and T.G.; formal analysis, N.K.; investigation, K.W. and T.G.; resources, N.K. and V.R.; data curation, N.K. and V.R.; writing—original draft preparation, N.K., V.R. and T.G.; writing—review and editing, K.W. and T.G.; visualization, T.G.; supervision, T.G.; project administration, K.W.; funding acquisition, K.W., N.K. and V.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Claas-Stiftung, grant number 6/2023/F.CLASS, as part of the project titled “Development of a technology for the management of waste biomass from agricultural production through methane fermentation and pellet production (incineration)”.

Data Availability Statement: All data and scripts supporting this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Detailed Predictions and Validation Results

Table A1. ANN LOSO prediction for Sugar beet pulp W1, source sheet Arkusz1 row 70, record XLSX_R070. This observation is excluded from training.

Output	Measured	Predicted	Absolute Error	APE (%)	sAPE (%)
CH4_fm	58.6131	58.7865	0.1734	0.30	0.30
Biogas_fm	107.0616	106.8662	0.1954	0.18	0.18
CH4_dm	323.6503	319.2726	4.3778	1.35	1.36
Biogas_dm	591.1738	598.8700	7.6962	1.30	1.29
CH4_odm	334.9724	325.2485	9.7239	2.90	2.95
Biogas_odm	611.8545	611.3413	0.5132	0.08	0.08
Mean across outputs	–	–	–	1.02	1.03

Table A2. ANN LOSO prediction for Top layer of beet pulp pile, source sheet Arkusz1 row 23, record XLSX_R023. This observation is excluded from training.

Output	Measured	Predicted	Absolute Error	APE (%)	sAPE (%)
CH4_fm	53.1750	55.8010	2.6260	4.94	4.82
Biogas_fm	103.9153	119.8830	15.9678	15.37	14.27
CH4_dm	370.1778	370.3328	0.1550	0.04	0.04
Biogas_dm	723.4058	703.0730	20.3328	2.81	2.85
CH4_odm	401.0539	420.7946	19.7406	4.92	4.80
Biogas_odm	783.7443	792.8562	9.1119	1.16	1.16
Mean across outputs	–	–	–	4.87	4.66

Table A3. ANN LOSO prediction for Turkey manure, source sheet Arkusz1 row 52, record XLSX_R052. This observation is excluded from training.

Output	Measured	Predicted	Absolute Error	APE (%)	sAPE (%)
CH ₄ _fm	87.6887	125.3481	37.6594	42.95	35.35
Biogas_fm	143.9648	192.3617	48.3969	33.62	28.78
CH ₄ _dm	171.3036	507.2565	335.9529	196.12	99.02
Biogas_dm	281.2412	722.7272	441.4860	156.98	87.95
CH ₄ _odm	200.2908	596.6704	396.3797	197.90	99.47
Biogas_odm	328.8316	853.6847	524.8531	159.61	88.77
Mean across outputs	–	–	–	131.20	73.22

Table A4. Secondary six-output LOSO summary and methane-FM detail. The first model triplet averages six outputs (secondary); the second reports methane-FM only. All 83 observations are included.

Label	n	ANN All	GP All	POD All	ANN CH ₄	GP CH ₄	POD CH ₄
Corn silage	3	72.52	26.22	46.08	74.15	36.88	52.80
Lettuce	3	15.37	17.26	21.23	17.88	22.11	37.65
Distillery stillage	3	107.34	104.15	31.62	183.50	160.52	13.45
Corn silage I	3	33.98	24.60	29.45	31.28	21.94	20.66
Corn silage II	3	54.36	34.29	39.01	49.51	24.65	32.49
Top layer of beet pulp pile	3	5.93	11.21	8.85	5.19	6.56	4.59
Center of the beet pulp pile	3	3.13	6.22	5.25	4.55	5.27	2.25
Corn silage III	3	42.08	25.60	36.36	37.22	22.61	35.86
Wheat straw SI	3	36.53	25.65	25.04	27.50	23.94	25.15
Wheat straw SII	3	14.28	11.41	6.98	19.27	12.07	11.23
Cattle manure I	3	28.39	11.88	14.79	64.08	13.86	20.72
Maize silage IV	3	120.93	48.29	57.42	66.42	38.33	49.93
Cattle manure II	3	12.63	13.53	13.33	20.96	17.20	16.74
Turkey manure	3	77.78	46.78	39.75	41.64	18.43	22.87
Cattle slurry I	3	85.40	95.29	75.86	165.81	200.00	200.00
Cattle slurry II	3	97.23	134.62	126.09	64.69	154.93	142.10
Chicken manure I	3	39.90	15.11	18.17	55.58	25.14	22.15
Chicken manure II	3	103.62	17.67	29.94	66.85	33.10	46.30
Sugar beet pulp W1	1	1.03	1.08	4.95	0.30	1.09	6.25
Sugar beet pulp W2	1	1.39	2.26	4.46	0.79	0.62	7.12
Sugar beet pulp W3	1	1.52	1.99	4.42	1.27	0.26	6.58
Pig slurry I	3	54.44	69.67	86.88	64.42	123.37	80.96
Stomach contents	3	90.53	87.74	80.23	119.56	97.50	80.67
Pig slurry II	3	118.13	112.96	114.02	183.07	180.47	176.95
Whey I	3	7.66	8.14	6.08	11.35	8.27	7.69
Whey II	2	10.12	10.62	12.46	12.19	22.60	31.31
Coconut flower after hydrolysis	3	53.85	41.22	37.89	56.42	43.32	41.93
Coconut flower	3	14.52	8.89	18.63	17.02	11.48	29.04
Chinese cabbage	3	41.85	75.80	74.58	83.98	173.26	133.21
Beet leaves	3	31.43	24.85	23.00	28.02	10.47	27.37

Table A5. ANN architecture diagnostic in five-fold CV, averaged over the two primary fresh-mass outputs. RMSE is in m³/Mg FM. Adam is used for both widths. The 20-neuron results do not replace the prespecified 10-neuron principal model.

Neurons	Training	RMSE	MAPE (%)	sMAPE (%)	R ²
10	Adam	25.81	73.78	40.19	0.899
20	Adam	24.30	41.36	32.99	0.912

Table A6. Per-output five-fold CV for GP. Predictions are pooled across held-out rows. GP intervals are nominal 95% observation intervals. The two FM rows are primary; normalized rows are secondary. Intervals are conditional on the fixed GP assumptions.

Output	RMSE	MAPE (%)	sMAPE (%)	R ²	PICP95 (%)	MPIW
CH4_fm	16.96	45.49	32.21	0.919	93.98	62.68
Biogas_fm	31.59	37.54	27.21	0.906	92.77	108.50
CH4_dm	47.98	19.43	16.53	0.877	84.34	143.68
Biogas_dm	83.73	16.77	15.19	0.853	81.93	230.03
CH4_odm	56.13	19.26	16.60	0.860	81.93	157.94
Biogas_odm	96.37	16.60	15.13	0.826	74.70	242.88

Table A7. Per-output five-fold CV for POD–RBF. Predictions are pooled across held-out rows. Each POD basis and rank are fitted on training rows only. The two FM rows are primary; normalized rows are secondary.

Output	RMSE	MAPE (%)	sMAPE (%)	R ²
CH4_fm	16.53	61.00	36.25	0.923
Biogas_fm	31.14	43.62	33.51	0.909
CH4_dm	40.58	16.84	14.54	0.912
Biogas_dm	73.67	14.29	13.38	0.886
CH4_odm	43.30	14.30	12.93	0.917
Biogas_odm	84.52	13.30	12.83	0.866

Table A8. Secondary six-output group-balanced LOSO sMAPE summaries and wins across 30 labels. IQR is the 75th minus 25th percentile, with linear interpolation.

Model	Mean (%)	Median (%)	IQR (%)	Wins
ANN	45.93	38.21	63.43	11
GP	37.17	24.73	36.66	10
POD–RBF	36.43	27.25	31.83	9

References

1. Witaszek, K.; Kupryaniuk, K.; Kupryaniuk, J.; Panasiewicz, J.; Czekala, W. Optimization of straw particle size for enhanced biogas production: A comparative study of wheat and rapeseed straw. *Energies* **2025**, *18*, 1794. [\[CrossRef\]](#)
2. Chen, J.W.; Chan, Y.J.; Arumugasamy, S.K.; Kazemi Yazdi, S. Process modelling and optimisation of methane yield from palm oil mill effluent using response surface methodology and artificial neural network. *J. Water Process Eng.* **2023**, *52*, 103493. [\[CrossRef\]](#)
3. Łukomska, A.; Witaszek, K.; Dach, J. Current state of development of demand-driven biogas plants in Poland. *Processes* **2025**, *13*, 2369. [\[CrossRef\]](#)
4. Labatut, R.A.; Angenent, L.T.; Scott, N.R. Biochemical methane potential and biodegradability of complex organic substrates. *Bioresour. Technol.* **2011**, *102*, 2255–2264. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Przybył, J.; Wojcieszak, D.; Garbowski, T. Effect of corn stover ensiling on methane production and carbon dioxide emissions. *Energies* **2024**, *17*, 6179. [\[CrossRef\]](#)
6. Angelidaki, I.; Alves, M.; Bolzonella, D.; Borzacconi, L.; Campos, J.L.; Guwy, A.J.; Kalyuzhnyi, S.; Jenicek, P.; van Lier, J.B. Defining the biomethane potential (BMP) of solid organic wastes and energy crops: A proposed protocol for batch assays. *Water Sci. Technol.* **2009**, *59*, 927–934. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Holliger, C.; Alves, M.; Andrade, D.; Angelidaki, I.; Astals, S.; Baier, U.; Bougrier, C.; Buffière, P.; Carballa, M.; de Wilde, V.; et al. Towards a standardization of biomethane potential tests. *Water Sci. Technol.* **2016**, *74*, 2515–2522. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Batstone, D.J.; Keller, J.; Angelidaki, I.; Kalyuzhnyi, S.V.; Pavlostathis, S.G.; Rozzi, A.; Sanders, W.T.M.; Siegrist, H.; Vavilin, V.A. The IWA Anaerobic Digestion Model No 1 (ADM1). *Water Sci. Technol.* **2002**, *45*, 65–73. [\[CrossRef\]](#)
9. Tisocco, S.; Weinrich, S.; Möller, H.B.; Ward, A.J.; Kilmartin, L.; Zhan, X.; Crosson, P. Machine learning vs. ADM1: Reliable biogas prediction with minimal data requirements in full-scale plants. *Environ. Sci. Ecotechnol.* **2026**, *29*, 100662. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Holubar, P.; Zani, L.; Hager, M.; Fröschl, W.; Radak, Z.; Braun, R. Advanced controlling of anaerobic digestion by means of hierarchical neural networks. *Water Res.* **2002**, *36*, 2582–2588. [\[CrossRef\]](#) [\[PubMed\]](#)

11. Zou, J.; Lü, F.; Chen, L.; Zhang, H.; He, P. Machine learning for enhancing prediction of biogas production and building a VFA/ALK soft sensor in full-scale dry anaerobic digestion of kitchen food waste. *J. Environ. Manag.* **2024**, *371*, 123190. [CrossRef] [PubMed]
12. Jeong, K.; Abbas, A.; Shin, J.; Son, M.; Kim, Y.M.; Cho, K.H. Prediction of biogas production in anaerobic co-digestion of organic wastes using deep learning models. *Water Res.* **2021**, *205*, 117697. [CrossRef] [PubMed]
13. Li, C.; He, P.; Peng, W.; Lü, F.; Du, R.; Zhang, H. Exploring available input variables for machine learning models to predict biogas production in industrial-scale biogas plants treating food waste. *J. Clean. Prod.* **2022**, *380*, 135074. [CrossRef]
14. Hunter, S.M.; Blanco, E.; Borrión, A. Predicting total biogas potential of food waste using the initial output of biogas potential tests as input data to train an artificial neural network. *Bioresour. Technol. Rep.* **2024**, *26*, 101845. [CrossRef]
15. Kowalczyk-Juśko, A.; Pochwatka, P.; Zaborowicz, M.; Czekala, W.; Mazurkiewicz, J.; Mazur, A.; Janczak, D.; Marczuk, A.; Dach, J. Energy value estimation of silages for substrate in biogas plants using an artificial neural network. *Energy* **2020**, *202*, 117729. [CrossRef]
16. Abdurrahman, A.; Sutiarto, L.; Ainuri, M.; Ushada, M.; Islam, M.P. A multilayer perceptron feedforward neural network and particle swarm optimization algorithm for optimizing biogas production. *Energies* **2025**, *18*, 1002. [CrossRef]
17. Tryhuba, I.; Tryhuba, A.; Hutsol, T.; Cieszeńska, A.; Andrushkiv, O.; Glowacki, S.; Bryś, A.; Slobodian, S.; Tulej, W.; Sojak, M. Prediction of biogas production volumes from household organic waste based on machine learning. *Energies* **2024**, *17*, 1786. [CrossRef]
18. Mukasine, A.; Sibomana, L.; Jayavel, K.; Nkurikiyeyezu, K.; Hitimana, E. Maximizing biogas yield using an optimized stacking ensemble machine learning approach. *Energies* **2024**, *17*, 364. [CrossRef]
19. Long, F.; Xu, M.; Liao, W.; Liu, H. Machine learning for predicting and optimizing the performance of a commercial-scale anaerobic digester with diverse feedstocks and operating conditions. *Bioresour. Technol.* **2025**, *435*, 132940. [CrossRef] [PubMed]
20. Meola, A.; Weinrich, S. Full-scale dynamic anaerobic digestion process simulation with machine and deep learning algorithms at intra-day resolution. *Appl. Energy* **2025**, *390*, 125781. [CrossRef]
21. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
22. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* **2018**, *180*, 68–77. [CrossRef] [PubMed]
23. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. [CrossRef] [PubMed]
24. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006. Available online: <https://gaussianprocess.org/gpml/> (accessed on 6 October 2026).
25. Adamski, M.; Czechowski, M.; Durczak, K.; Garbowski, T. Determination of the concentration of propionic acid in an aqueous solution by POD-GP model and spectroscopy. *Energies* **2021**, *14*, 8288. [CrossRef]
26. Garbowski, T.; Maier, G.; Novati, G. Diagnosis of concrete dams by flat-jack tests and inverse analyses based on proper orthogonal decomposition. *J. Mech. Mater. Struct.* **2011**, *6*, 181–202. [CrossRef]
27. Garbowski, T.; Maier, G.; Novati, G. On calibration of orthotropic elastic-plastic constitutive models for paper foils by biaxial tests and inverse analyses. *Struct. Multidiscip. Optim.* **2012**, *46*, 111–128. [CrossRef]
28. Kuleshov, V.; Fenner, N.; Ermon, S. Accurate uncertainties for deep learning using calibrated regression. *Proc. Mach. Learn. Res.* **2018**, *80*, 2796–2804.
29. Alhraishawi, A.; Abed Al Sailawi, A.S.; Hameed Al-Hamzawi, H.A. An explainable and uncertainty-aware machine-learning framework for cross-study prediction and optimization of methane yield in anaerobic digestion. *Bioresour. Technol. Rep.* **2026**, *35*, 102898. [CrossRef]
30. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
31. Saltelli, A.; Annoni, P.; Azzini, I.; Campolongo, F.; Ratto, M.; Tarantola, S. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Comput. Phys. Commun.* **2010**, *181*, 259–270. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.