Multi-Product Modeling of Consolidated Bioprocessing Using a Literature-Derived Dataset: A Multi-Output Learning Framework for Ethanol and Co-Products

Mark Korang Yeboah *, Ahmad Addo, Nana Yaw Asiedu

Posted Date: 30 March 2026

doi: 10.20944/preprints202603.2296.v1

Keywords: consolidated bioprocessing; literature-derived dataset; multi-product prediction; multioutput learning; missing-label handling; grouped cross-validation; random forest; biorefinery modeling

 doi:10.20944/preprints202603.2296.v1

Article

Multi-Product Modeling of Consolidated Bioprocessing Using a Literature-Derived Dataset: A Multi-Output Learning Framework for Ethanol and Co-Products

Mark Korang Yeboah [1,2,*], Ahmad Addo [2] and Nana Yaw Asiedu [2]

[1] Chair of Dynamics and Control, University of Duisburg–Essen, Lotharstraße, 47057 Duisburg, Germany
[2] Faculty of Mechanical and Chemical Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
* Correspondence: mark.korangyeboah@uni-due.de

Abstract

Consolidated bioprocessing (CBP) has been widely studied as an integrated route for converting biomass into biofuels and bioproducts, yet most quantitative modeling work has focused on ethanol as a single response. Because CBP systems can generate multiple products and co-products, this study develops a literature-derived benchmark for multi-product CBP modeling using a standardized dataset assembled from published endpoint experiments. Product prediction is formulated as both an observed-only product-wise problem and a joint multi-output problem, allowing direct comparison under study-aware grouped validation. The modeling space integrates biomass composition, pretreatment descriptors, microbial and consortium characteristics, reactor information, operating conditions, and engineered categorical descriptors of feedstock, pretreatment family, and process configuration. Predictive performance was strongly product-dependent and was shaped by target support and missing-label structure. The observed-only product-wise formulation consistently outperformed the joint missing-as-zero multi-output strategy, indicating that naive zero-filling of unreported products is not well suited to sparse literature-derived CBP data. Among the evaluated products, butanol showed the clearest predictive signal, ethanol was only moderately learnable, and the sparsest co-products remained too weakly supported for strong quantitative inference. Overall, the study provides a benchmark for multi-product CBP modeling and clarifies both the potential and the current limitations of literature-derived data for broader data-driven biorefinery analysis.

Keywords: consolidated bioprocessing; literature-derived dataset; multi-product prediction; multi-output learning; missing-label handling; grouped cross-validation; random forest; biorefinery modeling

1. Introduction

1.1. Background and Motivation

Consolidated bioprocessing (CBP) is widely regarded as a promising route for converting lignocellulosic biomass into fuels and chemicals because enzyme production, biomass hydrolysis, and fermentation can be integrated within a single process configuration [1–3]. By reducing dependence on externally supplied hydrolytic enzymes and simplifying process integration, CBP has been proposed as a strategy to lower operating costs and improve the feasibility of lignocellulosic biorefineries [3–5]. Recent advances in synthetic cellulosomes, engineered cellulolytic hosts, and artificial microbial consortia have further expanded the scope of CBP beyond proof-of-concept demonstrations toward more versatile biomass-conversion platforms [4–6].

At the same time, much of the CBP literature remains centered on bioethanol, reflecting ethanol's longstanding role in lignocellulosic biofuel development and its comparatively frequent reporting

 doi:10.20944/preprints202603.2296.v1

in experimental studies [2,7,8]. In parallel, the broader biorefinery concept increasingly emphasizes product diversification, in which biomass is converted not only to ethanol but also to value-added organic acids, higher alcohols, polymers, and other specialty products [2,9,10]. This shift motivates a broader analytical perspective on CBP systems, in which feedstock properties, pretreatment choices, microbial configurations, and reactor conditions are examined in relation to multiple possible outputs rather than to a single product alone.

1.2. From Ethanol-Centric Modeling to Multi-Product Analysis

Most experimental and modeling studies in CBP remain ethanol-centric. Recent examples include engineered cellulolytic or ethanologenic hosts, consortium-based ethanol production systems, and optimization-oriented studies focused primarily on ethanol endpoint responses [5–7,11,12]. This emphasis is understandable because ethanol remains both a major fuel target and one of the most frequently reported products in lignocellulosic fermentation studies [7,8].

Nevertheless, CBP is not limited to ethanol alone. Recent studies have reported CBP routes to butanol [13,14], malic acid [15], itaconic acid [16], and polyhydroxyalkanoates [17], while engineered and artificial microbial consortia have broadened the range of accessible conversion strategies and product outcomes [9,18]. These developments indicate that CBP can support a broader product portfolio, but they also expose a methodological challenge: the underlying data are fragmented across studies, product definitions, reporting conventions, feedstocks, and microbial systems. As a result, the field still lacks a unified empirical framework for examining how CBP design variables relate to multiple endpoint products.

1.3. Research Gap

Although machine learning, hybrid modeling, and data-driven process analytics are becoming more common in bioprocess engineering [19–23], their application to CBP remains comparatively limited and is still dominated by ethanol-focused studies or process-specific optimization tasks. Prior CBP modeling work has included mechanistic and cybernetic formulations, scale-up models, response-surface optimization, and single-product machine-learning studies [11,12]. However, literature-derived CBP data do not appear to have been systematically organized into a multi-product predictive framework that jointly examines ethanol and co-products within a unified endpoint-level dataset.

This gap is not only computational but also data-structural. Literature-derived CBP datasets are inherently heterogeneous, spanning different feedstocks, pretreatment methods, microbial strains or consortia, operating conditions, and output-reporting practices. Product coverage is also highly uneven: ethanol is reported frequently, whereas many co-products appear in only a relatively small number of studies. These characteristics make CBP a difficult but informative setting for evaluating multi-output learning, missing-label handling, and reference-aware validation. Addressing this gap can help clarify which products are meaningfully learnable from current literature-derived data and where predictive performance remains limited by sparse support, heterogeneous study design, and missing-label structure.

1.4. Objectives and Contributions

This study extends CBP modeling from an ethanol-centric perspective toward a literature-derived multi-product learning framework for ethanol and co-products. The work makes four main contributions.

First, a standardized CBP dataset is constructed and harmonized from literature sources, integrating endpoint records across multiple biomass types, pretreatment strategies, microbial systems, and reactor conditions into a common modeling structure.

Second, CBP product prediction is formulated as both a product-wise and a multi-output learning problem, enabling direct comparison between single-target and joint-target modeling strategies under the same reference-aware evaluation protocol.

 doi:10.20944/preprints202603.2296.v1

Third, the study assesses the extent to which shared CBP descriptors support prediction across ethanol and co-products, while identifying where predictive performance remains strongly product-dependent due to sparse support, heterogeneous study designs, and missing-label structure.

Fourth, the resulting framework is used as a benchmark to discuss the implications of data availability, feature engineering, validation design, and target sparsity for future CBP modeling and digital biorefinery analysis. In this sense, the present work is intended not as evidence of uniformly accurate multi-product prediction but as a data-centric and methodological assessment of how far current literature-derived CBP data can support robust multi-product learning.

2. Dataset Construction and Problem Formulation

2.1. Literature-Derived Dataset

A literature-derived CBP dataset was assembled at the experimental-endpoint level from published studies. The final dataset comprised 640 records and 118 variables after removal of near-constant binary descriptors. Each record represented a single experimental endpoint and included a study-level grouping identifier, along with feedstock, pretreatment, microbial system, reactor, and operating condition descriptors, as well as one or more reported product titers. An endpoint-level representation was adopted because it supports direct comparison of ethanol and co-products within a common supervised-learning framework while preserving traceability to the source literature. Because the literature sources reported product measurements with inconsistent sampling schedules and uneven temporal detail, the dataset was harmonized at the experimental-endpoint level rather than as a unified time-resolved panel. Additional details on the supplementary Excel workbook and dataset description are provided in the Supplementary Information.

2.2. Selection of Products and Output Harmonization

Eight endpoint products were considered as supervised targets: ethanol, acetate, lactate, formate, butanol, succinic acid, *D*-glucaric acid, and hydrogen. Reported product names and alternative spellings were harmonized into a unified set of response variables. Liquid-phase products were standardized to $\mathrm{g\,L^{-1}}$, whereas hydrogen was represented in $\mathrm{mmol\,L^{-1}}$. Product support was highly unbalanced across the dataset, with ethanol reported most frequently and several co-products available only for a limited subset of records. This imbalance motivated evaluating both joint and product-specific prediction settings. A detailed description of product harmonization, response units, and per-product coverage is provided in the Supplementary Information.

2.3. Input Variables and Feature Groups

Predictor variables were organized into four groups reflecting the CBP workflow. The first group described biomass and feedstock characteristics, including source category and compositional attributes such as cellulose, hemicellulose, and lignin content. The second group captured pretreatment conditions, including reagent information, temperature, residence time, pH, washing, detoxification, and severity-related descriptors. The third group represented the microbial system, including inoculum level, consortium structure, and indicators of microbial composition. The fourth group described reactor and operating conditions, including working volume, substrate loading, temperature, pH, residence time, agitation, and scale-related descriptors.

2.4. Feature Engineering

Feature engineering was used to translate heterogeneous literature annotations into modeling-ready descriptors while retaining process interpretability. Derived variables captured feedstock class, pretreatment family, principal reagent category, microbial-system configuration, kingdom composition, reactor scale, microorganism presence, and informative missingness in partially reported pretreatment fields. Near-constant binary descriptors were removed to reduce redundancy and improve model stability.

 doi:10.20944/preprints202603.2296.v1

2.5. Missing Data Handling and Data Preprocessing

All preprocessing steps were performed independently within each training fold to avoid information leakage. Numeric predictors were imputed using the median and scaled, whereas categorical predictors were imputed using the most frequent category and encoded for model input. Features with zero variance after preprocessing were removed.

Missing response values were addressed using two complementary formulations. In the primary *observed-only product-wise formulation*, a separate model was trained for each product using only records in which that product was reported. This formulation avoided imposing assumptions on unreported outcomes. In the secondary *joint zero-filled multi-output formulation*, missing product values in the training data were provisionally set to zero for model fitting, whereas evaluation remained restricted to observed validation targets. The second formulation was included as a sensitivity analysis for sparse and uneven product reporting.

Outlier handling was also carried out within training folds only. Eligible numeric predictors were screened using an interquartile-range rule, and records with the greatest aggregate predictor-level outlier burden were removed subject to a maximum trimming fraction of 30%. Validation records were not trimmed.

2.6. Problem Definition

The primary task was endpoint-level prediction of multiple CBP products from feedstock, pretreatment, microbial, and reactor descriptors. Let $\mathbf{x}_i \in \mathbb{R}^p$ denote the predictor vector for experiment i, and let $\mathbf{y}_i \in \mathbb{R}^8$ denote the corresponding vector of harmonized product responses. Two prediction settings were examined: an *observed-only product-wise formulation*, in which separate models were trained for individual products using only observed labels, and a *joint zero-filled multi-output formulation*, in which a single model predicted the full product vector after zero-filling missing training targets. This dual formulation enabled direct comparison between target-specific learning and shared-output learning under sparse literature-derived labeling.

3. Modeling Methodology

3.1. Observed-Only Product-Wise Formulation

Under the observed-only product-wise formulation, a separate model was trained for each product using only records with observed values for that product. This formulation avoided imposing assumptions on unreported outcomes and served as the primary analytical basis. The benchmark model set was intentionally compact and comprised a mean-response baseline, ridge regression, partial least squares (PLS) regression, random forest, and extremely randomized trees. These models were selected to span a range of inductive biases, including constant prediction, linear shrinkage, latent-variable regression, and nonlinear tree-based ensembling. A compact benchmark set was preferred to support stable comparisons under nested cross-validation on a sparse, heterogeneous literature-derived dataset.

3.2. Joint Zero-Filled Multi-Output Formulation

Under the joint zero-filled multi-output formulation, a single model was trained to predict the full product vector simultaneously. Because literature-derived CBP studies often report only a subset of products, missing target values in the training data were provisionally set to zero for model fitting, whereas validation metrics were computed only for targets explicitly reported in the corresponding validation records. This formulation was included as a secondary sensitivity analysis to assess whether shared process descriptors could improve prediction across related products. However, because an unreported product does not necessarily indicate true zero formation, this formulation was not treated as the primary analytical basis.

 doi:10.20944/preprints202603.2296.v1

3.3. Training and Hyperparameter Tuning

Model development followed a nested cross-validation design in which hyperparameters were selected within inner folds and performance was assessed only in outer folds. The tuning space was intentionally modest. For ridge regression, the regularization parameter α was varied over $\{1, 10, 50\}$. For PLS regression, the number of latent components was varied over $\{2, 4, 6\}$. For random forest, the number of trees was fixed at 300 while tree depth and minimum leaf size were varied. For extremely randomized trees, the number of trees was fixed at 400 while tree depth and minimum leaf size were varied. Inner-loop model selection was based on mean macro-averaged RMSE. For PLS regression, the effective number of latent components was additionally constrained by the rank and sample support available within each training fold. After inner-loop selection, the chosen model was refit on the full outer-training fold and used to generate outer-fold predictions.

3.4. Grouped Validation Strategy

To reduce optimistic bias arising from related experiments reported within the same source study, all data splitting was performed with study-aware grouping. The outer loop used five grouped folds and the inner loop used three grouped folds, subject to the number of distinct studies available. This design ensured that records from the same source were not split across training and validation sets within a given evaluation round. Group-aware validation was especially important because multiple records could share study-specific context, including feedstock selection, microbial design, and experimental practice.

3.5. Performance Metrics

Predictive performance was evaluated using root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and Spearman rank correlation. All metrics were computed from outer-fold predictions only. Under the joint zero-filled multi-output formulation, metrics were calculated separately for each product using only observed validation entries and then summarized using macro-averaged values across targets. RMSE and MAE were emphasized because they preserve the physical units of the response variables, whereas R^2 and Spearman correlation were included to assess explained variance and rank-order consistency under pronounced target imbalance.

3.6. Model Interpretation

Model interpretation was handled cautiously and focused on relative variable importance and the stability of model behavior across products and validation folds. Interpretation was used to assess whether certain descriptor groups showed consistent associations with predictive performance across targets, while avoiding strong mechanistic claims for products with limited statistical support. Given the heterogeneity of the source literature and the sparse support for several co-products, all interpretation was treated as association-based rather than causal.

The modeling pipeline comprised dataset harmonization, fold-wise preprocessing with training-only trimming, comparison of the observed-only product-wise formulation and the joint zero-filled multi-output formulation, and shared benchmarking under grouped nested cross-validation, as shown in Figure 1.

 doi:10.20944/preprints202603.2296.v1

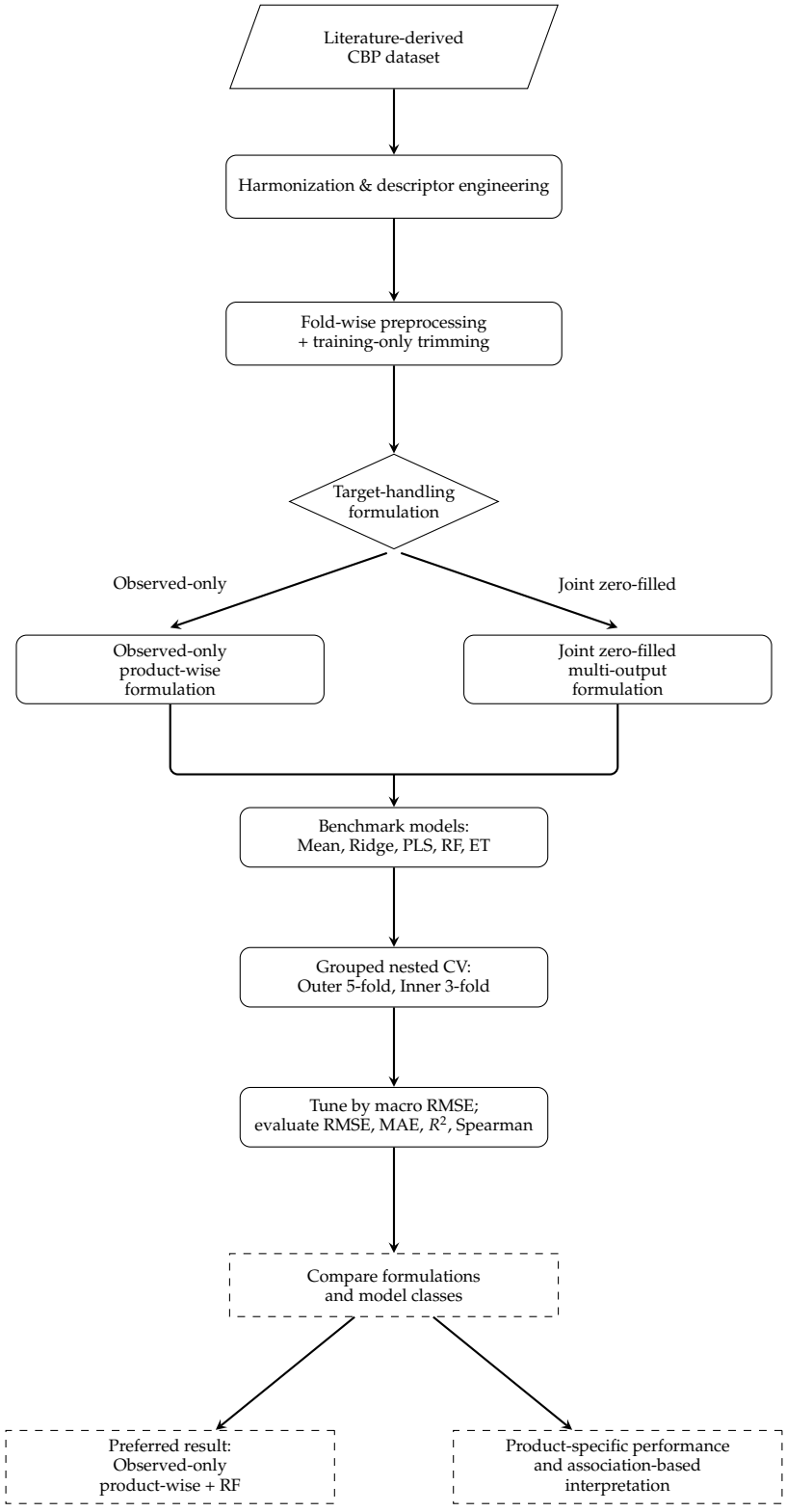


Figure 1. Simplified methodological workflow for literature-derived multi-product CBP modeling. The pipeline branches only at the target-handling step, contrasting the observed-only product-wise formulation with the joint zero-filled multi-output formulation, while preprocessing, benchmarking, grouped nested cross-validation, and evaluation remain shared across formulations.

3.7. Computational Reproducibility

The analyses were implemented in Python 3.13.5 and developed in Visual Studio Code, with a fixed random seed of 42. The workflow used numpy 2.3.2, pandas 2.3.1, matplotlib 3.10.3, scipy 1.16.0,

 doi:10.20944/preprints202603.2296.v1

scikit-learn 1.7.1, and joblib 1.5.1. The code and processed dataset will be made publicly available upon acceptance.

4. Results and Discussion

4.1. Dataset Support and Missingness

As summarized in Table 1, the literature-derived CBP dataset comprised 640 experimental endpoint records described by 118 variables and eight supervised product targets. The dataset integrates mixed biomass, pretreatment, microbial-system, reactor, and operating descriptors within a common endpoint-level representation. This unified structure enables joint benchmarking across products, but it also introduces substantial heterogeneity that must be considered when interpreting predictive performance.

Target support was highly uneven across products, as shown in Figure 2. Ethanol was the only well-supported response, with 543 observed records out of 640 total entries (84.8%). All other products were much sparser, including acetate (147, 23.0%), butanol (46, 7.2%), lactate (41, 6.4%), formate (26, 4.1%), hydrogen (24, 3.8%), succinic acid (12, 1.9%), and *D*-glucaric acid (10, 1.6%). This imbalance is a defining feature of the dataset and indicates that predictive learnability is likely to vary strongly by product rather than remain uniform across outputs.

For most co-products, unreported outcomes far outnumber observed outcomes, limiting statistical support and increasing sensitivity to missing-label assumptions. The product-level results that follow should therefore be interpreted primarily as evidence of differential learnability in a sparse, heterogeneous literature-derived dataset rather than as evidence of uniformly reliable multi-product prediction. A complete per-product coverage summary and additional workbook details are provided in the Supplementary Information.

Table 1. Short summary of the literature-derived CBP dataset used in this study.

Attribute	Value
Records	640
Variables	118
Representation	Experimental endpoint level
Supervised targets	8 products
Input descriptor groups	Biomass, pretreatment, microbial system, reactor/operation
Response standardization	Liquid products in g L^{-1}; hydrogen in mmol L^{-1}

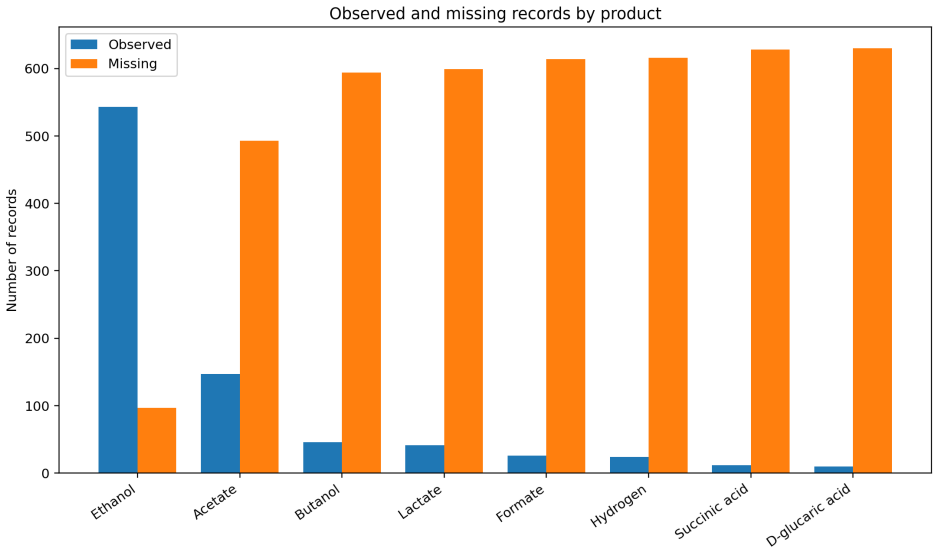


Figure 2. Observed and unreported records by product in the literature-derived CBP dataset. Ethanol was the only well-supported response, whereas all co-products were much more sparsely reported. This imbalance is a defining feature of the dataset and constrains model training, validation, and interpretation.

 doi:10.20944/preprints202603.2296.v1

4.2. Final Model Selection and Tuning

Across both target-handling formulations, Random Forest was selected as the final model with the same regularized configuration: 300 trees, maximum depth of 8, and minimum leaf size of 3 (Table 2). This convergence indicates that the main difference between formulations did not arise from a different preferred model class, but from how missing response values were handled.

Inner cross-validation performance supports this interpretation. The observed-only product-wise formulation achieved a lower mean inner macro RMSE (7.27) than the joint zero-filled multi-output formulation (10.69), despite selecting the same model family and hyperparameter configuration. Thus, the two formulations converged on the same model class, but not on the same level of tuning performance.

Table 2. Selected Random Forest configuration and best inner-cross-validation performance for each formulation.

Formulation	Trees	Max depth	Min leaf	Mean inner macro RMSE
Joint zero-filled multi-output	300	8	3	10.688
Observed-only product-wise	300	8	3	7.266

4.3. Comparison of Target-Handling Strategies

The two target-handling formulations were compared using nested grouped cross-validation. Under the joint zero-filled multi-output formulation, missing product values were set to zero during training and evaluation was restricted to observed validation targets. Under the observed-only product-wise formulation, a separate model was trained for each product using only records with observed labels. In both settings, Random Forest was selected as the best overall model, but the observed-only product-wise formulation consistently yielded stronger out-of-fold performance, as summarized in Table 3.

The observed-only product-wise formulation reduced macro RMSE from 12.68 to 10.49 and macro MAE from 9.40 to 6.16 relative to the joint zero-filled multi-output formulation. Macro R^2 improved from -4.29 to -0.04, and macro Spearman correlation increased from approximately zero to 0.255. Although absolute performance remained imperfect, the direction and consistency of these changes indicate that zero-filling missing targets degraded both calibration and rank preservation, as summarized in Table 3.

The broader model-by-formulation comparison supports the same conclusion. Across candidate models, the observed-only product-wise formulation generally achieved lower macro RMSE, whereas macro R^2 remained weak or unstable in several settings despite the relative improvement over the joint zero-filled multi-output formulation, as shown in Figures 3 and 4. These results indicate that target-handling is a consequential modeling choice in this sparse literature-derived dataset rather than a minor preprocessing detail.

Table 3. Best-model out-of-fold macro metrics by target-handling formulation. In both cases, Random Forest was selected as the best overall model.

Formulation	Macro RMSE	Macro MAE	Macro R^2	Macro Spearman
Joint zero-filled multi-output	12.68	9.40	-4.29	-0.003
Observed-only product-wise	10.49	6.16	-0.04	0.255

 doi:10.20944/preprints202603.2296.v1

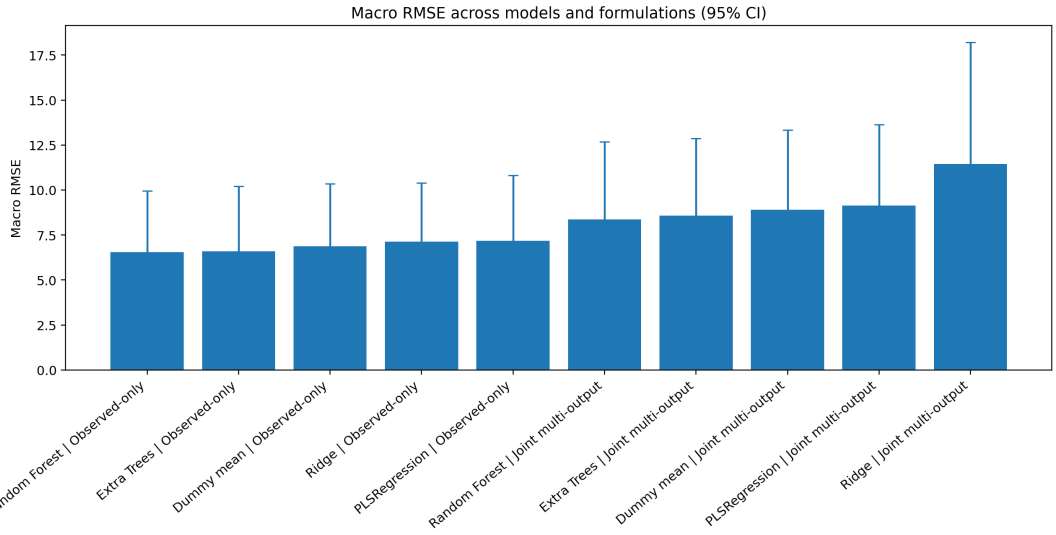


Figure 3. Nested cross-validation macro RMSE across candidate models and target-handling formulations (95% confidence intervals). The observed-only product-wise formulation generally yielded lower error than the joint zero-filled multi-output formulation.

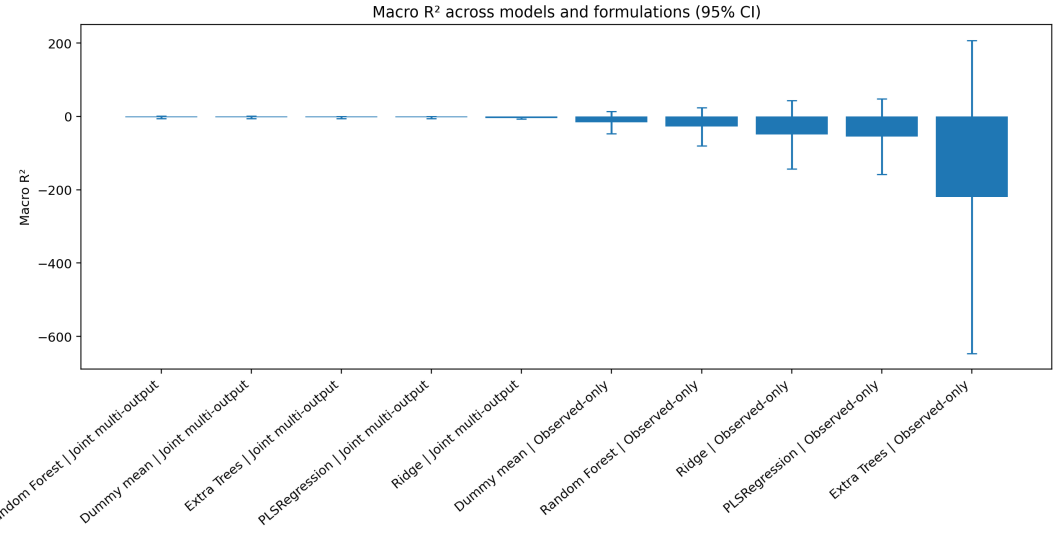


Figure 4. Nested cross-validation macro R^2 across candidate models and target-handling formulations (95% confidence intervals). Although the observed-only product-wise formulation improved relative performance, R^2 remained weak or unstable in several model settings.

4.4. Outer-Fold Model Ranking and Selection Stability Under the Preferred Formulation

Because the observed-only product-wise formulation emerged as the preferred approach, the relative performance of candidate models within this setting is informative. Averaged across outer folds, Random Forest achieved the lowest macro RMSE (6.54), followed closely by Extra Trees (6.59), the mean-response baseline (6.87), ridge regression (7.13), and partial least squares regression (7.18), as shown in Table 4. This ranking indicates that tree-based ensembles were the most competitive model class under the preferred formulation, although the margin over the simplest baseline was modest.

The outer-fold selection frequencies provide a complementary view of the same result. Under the preferred observed-only product-wise formulation, Random Forest ranked first in three of five outer folds, while Extra Trees ranked first in the remaining two. Under the joint zero-filled multi-output formulation, Random Forest ranked first in all five outer folds, and no linear model or mean-response baseline ranked first in either formulation, as summarized in Table 5. Taken together, these results

 doi:10.20944/preprints202603.2296.v1

indicate a stable preference for nonlinear ensemble methods across folds rather than a conclusion driven by a single favorable split.

Random Forest was therefore retained as the main reference model for the subsequent product-specific analysis. Although Extra Trees was numerically close, Random Forest provided the strongest overall balance in the aggregate comparison while remaining the most frequently selected model under the preferred formulation.

Table 4. Mean outer-fold macro performance of candidate models under the preferred observed-only product-wise formulation.

Model	Mean RMSE	SD	Mean MAE	Spearman
Random Forest	6.54	3.88	4.64	0.250
Extra Trees	6.59	4.11	4.75	0.250
Mean baseline	6.87	3.97	5.07	—
Ridge	7.13	3.72	5.40	0.040
PLS regression	7.18	4.16	5.44	0.078

Table 5. Outer-fold best-model frequency under each target-handling formulation.

Model	Joint zero-filled multi-output	Observed-only product-wise
Random Forest	5/5	3/5
Extra Trees	0/5	2/5
Ridge	0/5	0/5
PLS regression	0/5	0/5
Mean baseline	0/5	0/5

4.5. Effect of Fold-Wise Trimming

Trimming was learned exclusively from the training folds, thereby avoiding leakage into validation data. Even with this precaution, the trimming burden differed substantially between the two formulations. The joint zero-filled multi-output formulation consistently required heavier row removal than the observed-only product-wise formulation, and in the final fitted model it reached the prespecified trimming ceiling of 30%. This is important for interpretation because the joint zero-filled multi-output formulation not only performed worse, but also depended on more aggressive filtering to do so, as summarized in Table 6 and Figure 5.

Table 6. Summary of trimming burden for the final selected formulations.

Formulation	Mean fold-wise trimming burden (%)	Final-fit trimmed fraction (%)
Joint zero-filled multi-output	25.7	30.0
Observed-only product-wise	12.5	—

 doi:10.20944/preprints202603.2296.v1

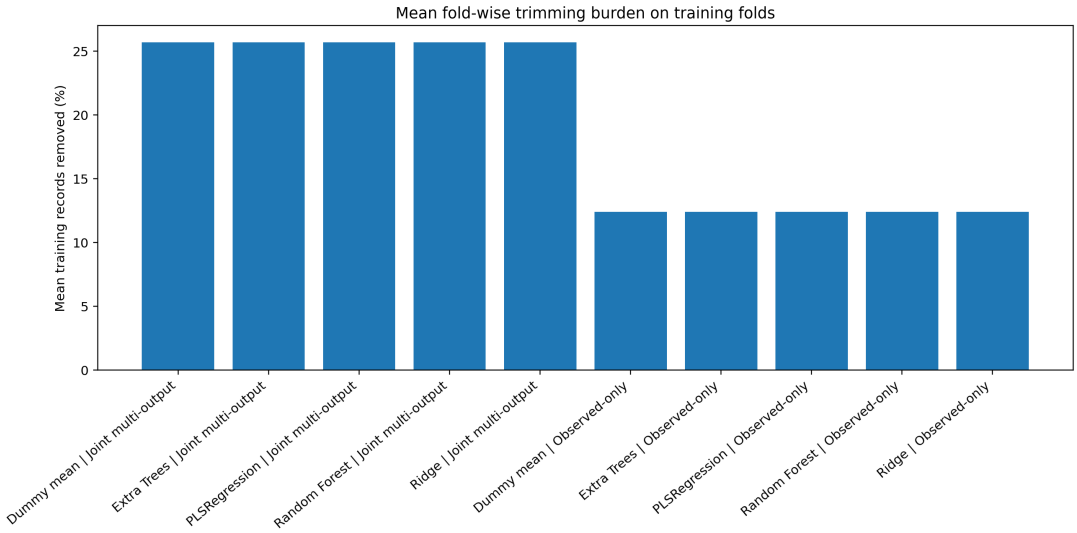


Figure 5. Mean fold-wise trimming burden on training folds. The joint zero-filled multi-output formulation consistently required heavier filtering than the observed-only product-wise formulation.

The product-specific trimming record under the preferred observed-only product-wise formulation provides additional context. In the final target-wise fit, trimming removed 29.8% of ethanol rows, 19.7% of acetate rows, 29.2% of hydrogen rows, and 20.0% of *D*-glucaric acid rows, but only 2.2% of butanol rows and none of the formate or succinic acid rows. This uneven burden suggests that part of the product-specific variation in predictive performance reflects differences in data regularity as well as differences in raw target support, as shown in Table 7.

Table 7. Product-specific row support and trimming burden for the final observed-only product-wise Random Forest fit.

Product	Final observed training rows	Trimmed rows	Trimmed fraction (%)
Ethanol	543	162	29.8
Acetate	147	29	19.7
Butanol	46	1	2.2
Lactate	41	2	4.9
Formate	26	0	0.0
Hydrogen	24	7	29.2
Succinic acid	12	0	0.0
D-glucaric acid	10	2	20.0

4.6. Product-Specific Performance and Support-Dependent Learnability

Predictive performance was strongly product-dependent. Under the observed-only product-wise formulation, butanol showed the clearest evidence of learnability, with RMSE decreasing from 7.17 to 5.92 and Spearman correlation increasing from 0.220 to 0.612 relative to the joint zero-filled multi-output formulation. Ethanol also improved, with RMSE decreasing from 15.42 to 14.82 and Spearman correlation increasing from 0.390 to 0.476. Acetate changed only marginally between formulations, lactate remained weak in both settings, and hydrogen improved from an RMSE of 39.18 to 29.83 but remained weak overall. The sparsest products, including succinic acid and *D*-glucaric acid, remained too data-limited to support robust quantitative conclusions, as shown in Figure 6.

These results indicate that the current literature-derived dataset supports learnability only for a subset of products, rather than providing uniformly accurate multi-product regression. Product support was clearly important, but it was not sufficient on its own to guarantee a strong predictive structure. Under the preferred observed-only product-wise formulation, ethanol remained the most frequently reported product ($n = 543$) yet showed only moderate rank recovery (Spearman = 0.476),

 doi:10.20944/preprints202603.2296.v1

whereas butanol, despite much lower support ($n = 46$), showed the strongest rank agreement (Spearman = 0.612). By contrast, acetate, lactate, and hydrogen remained weakly recoverable under the same workflow, as summarized in Table 8.

The comparison also suggests that products differed in their recoverable signal under the current feature space. Ethanol likely benefited from high support but remained heterogeneous because it aggregated diverse feedstocks, pretreatments, and microbial systems. Butanol appeared more recoverable relative to its sample size, possibly reflecting a more coherent subset of study designs or more structured process–response behavior. In contrast, acetate and hydrogen showed weak calibration and unstable rank ordering, while lactate showed low absolute error largely because of a narrow response range rather than strong predictive recovery. Detailed per-product values are provided in Supplementary Table S5.

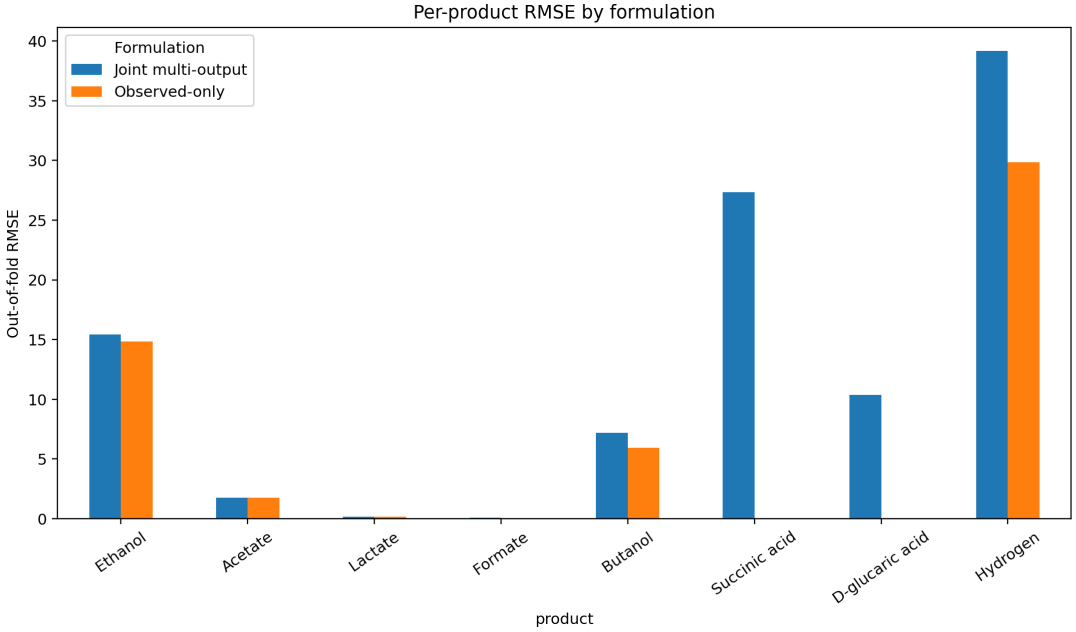


Figure 6. Per-product RMSE under the two target-handling formulations. The observed-only product-wise formulation shows the clearest gains for butanol, hydrogen, and ethanol, whereas the sparsest products remain weak overall.

Table 8. Observed support and best-model out-of-fold performance under the preferred observed-only product-wise formulation.

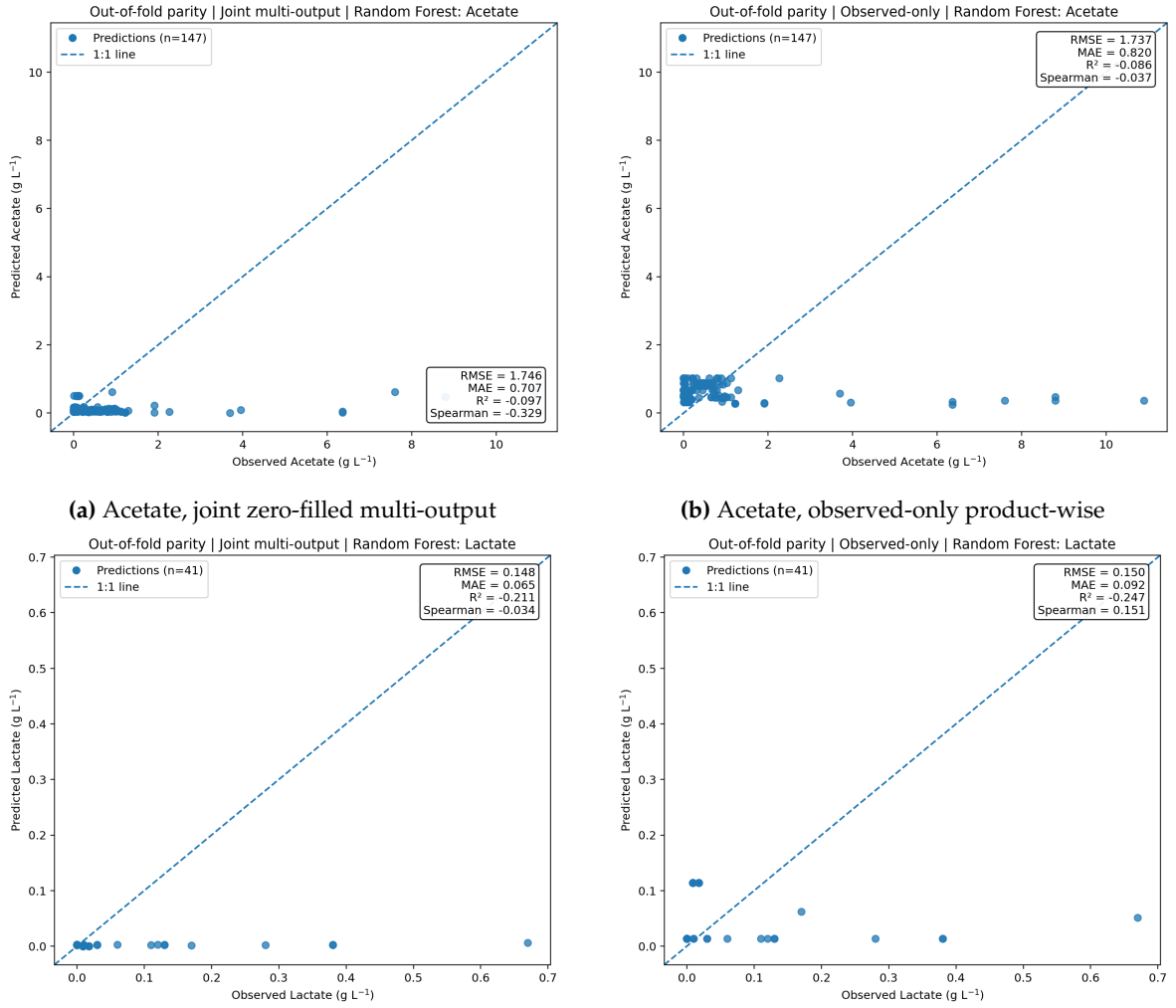
Product	Observed records	RMSE	Spearman
Ethanol	543	14.82	0.476
Acetate	147	1.74	-0.037
Butanol	46	5.92	0.612
Lactate	41	0.150	0.151
Hydrogen	24	29.83	0.074

4.7. Diagnostic Error Structure for Moderate-to-Weakly Learnable Products

The parity and residual diagnostics for acetate and lactate indicate that the main limitation of the current framework is not purely random error, but systematic compression toward lower predicted values. For acetate, the two formulations produced nearly identical RMSE values, yet both parity plots show substantial deviation from the 1:1 line at higher observed concentrations, and the residual plots show increasingly negative errors as the observed value increases. This pattern indicates persistent underprediction of larger acetate responses, even under the observed-only product-wise formulation. The slight improvement in rank agreement under the observed-only product-wise

 doi:10.20944/preprints202603.2296.v1

formulation therefore appears to reflect only a modest gain in relative ordering rather than a substantial improvement in calibration, as shown in Figures 7 and 8.



(a) Acetate, joint zero-filled multi-output

(b) Acetate, observed-only product-wise

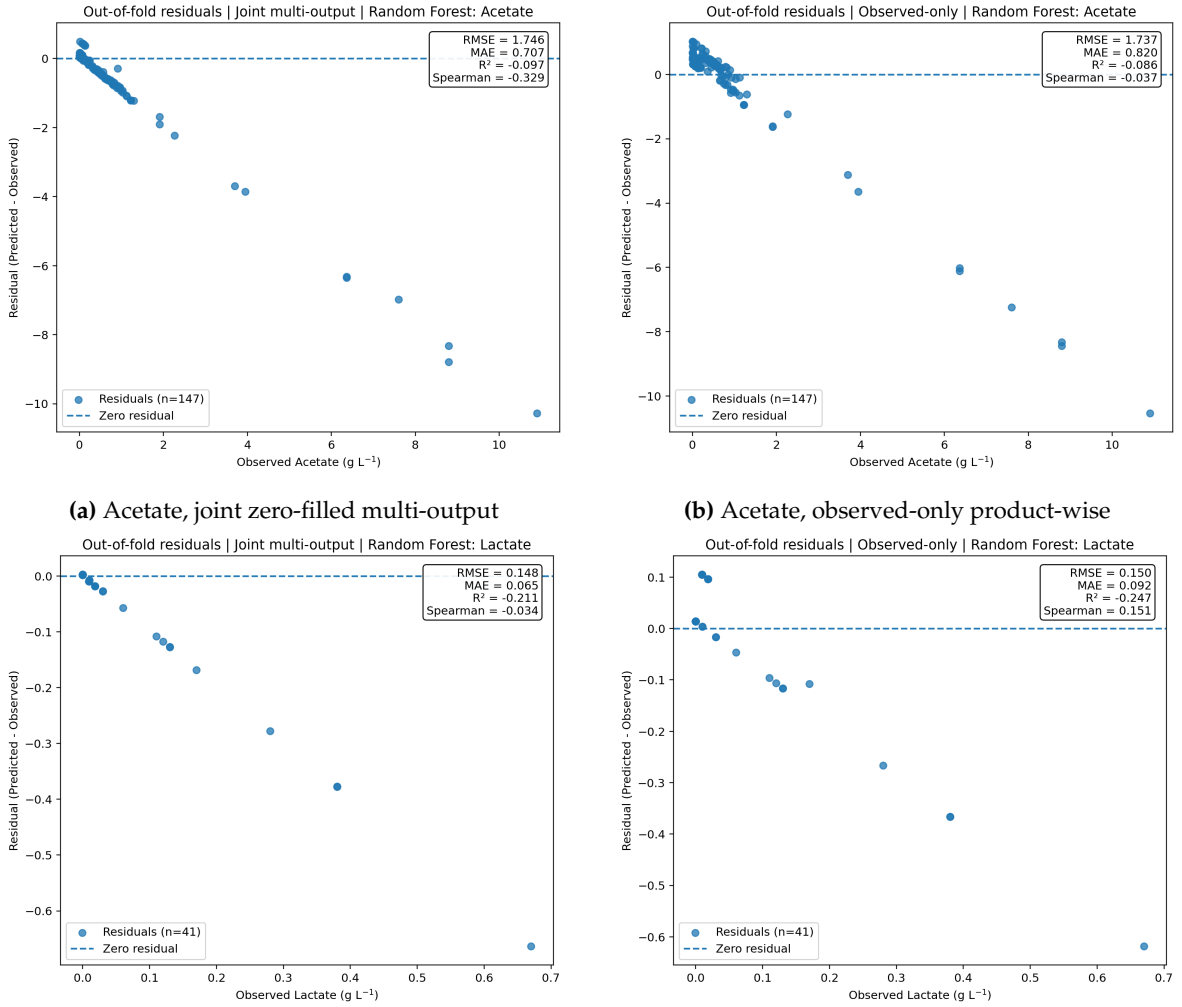
(c) Lactate, joint zero-filled multi-output

(d) Lactate, observed-only product-wise

Figure 7. Parity diagnostics for acetate and lactate under the two target-handling formulations. For both products, predictions remain compressed toward low values, and departures from the 1:1 line become more evident at higher observed responses. The observed-only product-wise formulation provides only limited improvement in calibration.

Lactate showed a related but even more weakly structured pattern. Absolute RMSE was small because the response range was narrow, but both parity and residual diagnostics indicated that larger observed values were still poorly recovered. The observed-only product-wise formulation modestly improved Spearman correlation relative to the joint zero-filled multi-output formulation, yet the overall fit remained weak and the residuals remained predominantly negative at the upper end of the observed range. In practical terms, these plots show that low absolute error for lactate should not be overinterpreted as evidence of robust predictability, as shown in Figures 7 and 8.

 doi:10.20944/preprints202603.2296.v1



(a) Acetate, joint zero-filled multi-output **(b)** Acetate, observed-only product-wise

(c) Lactate, joint zero-filled multi-output **(d)** Lactate, observed-only product-wise

Figure 8. Residual diagnostics for acetate and lactate under the two target-handling formulations. Both products show increasingly negative residuals at higher observed values, indicating persistent underprediction of upper-range responses. The observed-only product-wise formulation reduces this bias only modestly.

Taken together, these diagnostics indicate that the observed-only product-wise formulation is the more defensible approach for sparse literature-derived CBP data. At the same time, they show that product-specific bias remains substantial even for responses with relatively favorable error magnitudes. The current framework is therefore best interpreted as a benchmark of product-dependent learnability rather than as a uniformly accurate predictive platform across CBP products, as shown in Figures 7 and 8.

4.8. Limitations

Several limitations shape the interpretation of the present results. First, the dataset was assembled from independent CBP studies and therefore inherits substantial cross-study heterogeneity in feedstock characterization, pretreatment descriptions, microbial-system reporting, reactor operation, sampling practices, and analytical measurements. Such variability is characteristic of the current CBP literature and cannot be removed fully through endpoint-level harmonization alone [1–3,5]. Consequently, part of the observed prediction error likely reflects differences in study design and reporting quality rather than only biological or process-level behavior.

Second, target support was highly uneven across products. Ethanol was comparatively well represented, whereas several co-products were available for only a small subset of records. This imbalance limited statistical support for sparse products and contributed to strongly product-dependent

 doi:10.20944/preprints202603.2296.v1

predictive performance. The resulting framework is therefore more defensible as a benchmark of data-supported learnability than as a uniformly accurate predictive model for all CBP outputs.

Third, missing labels remain intrinsically ambiguous in literature-derived CBP data. An unreported product may reflect true absence, concentration below detection, selective reporting, or omission from the experimental objective. Although the observed-only product-wise formulation proved more reliable than the joint zero-filled multi-output formulation for this dataset, neither strategy fully resolves the underlying missingness mechanism. This issue is especially important in heterogeneous bioprocess datasets, where reporting practices can vary substantially across studies [19,20,22].

Fourth, the analysis was limited to endpoint prediction. CBP systems are inherently dynamic, and endpoint titers do not capture temporal interactions among hydrolysis, growth, intermediate accumulation, and product conversion. As a result, the present models cannot resolve kinetic behavior, transient pathway competition, or timing-dependent trade-offs that are central to many CBP systems [4,6,18].

Finally, the benchmark was intentionally conservative. Grouped nested cross-validation and fold-wise trimming reduced optimistic bias, but the final comparison still relied on tabular descriptors and relatively robust model classes suited to sparse, mixed-type data. The resulting interpretation remains association-based rather than causal, and the identified variables should not be interpreted as definitive process drivers [19,20,23].

4.9. Implications and Future Directions

Taken together, the results support interpreting this study as a benchmark and methodological contribution rather than as evidence of uniformly accurate multi-product prediction across all CBP outputs. The unified literature-derived framework proved feasible, but predictive signal remained uneven across products and was strongly shaped by target support, missingness, and cross-study heterogeneity. Among the evaluated products, butanol showed the clearest evidence of learnability, ethanol showed only moderate structure, and the sparsest co-products remained weakly constrained. The observed-only product-wise formulation consistently outperformed the joint zero-filled multi-output formulation and required less trimming, indicating that, for literature-derived CBP datasets with substantial label sparsity, treating unreported products as missing is more defensible than assigning them a value of zero.

The main contribution is therefore methodological as much as predictive. The study establishes a grouped nested cross-validation workflow for multi-product CBP analysis, shows that missing-label assumptions materially affect model behavior, and highlights product support as a primary determinant of learnability in literature-derived bioprocess data. The results also indicate that learnability is product-dependent and, to some extent, model-class-dependent.

Future progress in multi-product CBP modeling will depend first on stronger data resources. Expansion of the literature-derived dataset should be accompanied by tighter unit harmonization, clearer handling of detection limits, and richer metadata for feedstock composition, pretreatment severity, microbial configuration, reactor design, and analytical method. More consistent reporting of co-products and operating conditions across CBP studies would improve comparability and reduce ambiguity in downstream modeling [1,2,10].

A second priority is more explicit treatment of sparse and partially observed targets. Two-stage formulations that separate product occurrence from product magnitude, as well as censored, zero-inflated, masked multitask, or hierarchical models, may be better suited to literature-derived CBP data than a single regression formulation applied uniformly across all products [19–22].

A third direction is the integration of temporal and mechanistic structure. Time-resolved CBP datasets would enable prediction of product trajectories rather than final titers alone, while hybrid approaches could combine process knowledge with machine learning to represent coupled phenomena such as hydrolysis, uptake, and by-product formation more realistically [4,6,18,23].

Model interpretation and validation should also be strengthened. Future studies should include target-specific dependence analysis, local explanation methods, uncertainty-aware attribution, and

 doi:10.20944/preprints202603.2296.v1

stratified error analysis across feedstock classes, pretreatment families, and microbial-system types. External validation using source-wise holdout sets, temporal splits, and independent datasets will also be needed to assess transportability beyond the present literature pool [10,19,20].

The long-term objective is not only to improve predictive accuracy, but also to determine which CBP products are reliably learnable from current evidence, which additional descriptors most improve generalization, and how data-centric design can better support biorefinery-oriented process development [9,10].

5. Summary and Conclusions

This study developed a literature-derived modeling framework for consolidated bioprocessing (CBP) that extends analysis beyond ethanol to a multi-product setting. A standardized endpoint-level dataset was assembled from heterogeneous CBP studies, and grouped validation was used to reduce leakage across related experiments while accommodating mixed numerical and categorical descriptors. Product prediction was evaluated under two target-handling formulations to assess how missing-label assumptions affect model behavior.

Across the evaluated models, Random Forest provided the strongest overall balance of performance and robustness. The observed-only product-wise formulation consistently outperformed the joint zero-filled multi-output formulation and generally required less aggressive trimming. These results indicate that missing-label treatment is a central modeling decision in literature-derived CBP data.

Predictive performance remained strongly product-dependent. Butanol showed the clearest evidence of learnability, ethanol showed only moderate predictive structure, and several sparsely reported co-products remained too weakly supported for robust quantitative conclusions. The results therefore indicate both the feasibility and the present limits of multi-product CBP learning: a unified benchmarking framework can be constructed, but recoverable signal depends strongly on target support, missingness structure, and cross-study heterogeneity.

The principal contribution of the study is methodological as much as predictive. The framework establishes a grouped nested cross-validation workflow for multi-product CBP analysis, shows that missing-target assumptions materially affect model conclusions, and highlights product support as a primary determinant of learnability in literature-derived bioprocess datasets. The study should therefore be interpreted as a benchmark for multi-product CBP modeling rather than as evidence of uniformly accurate prediction across all outputs.

Author Contributions: Conceptualization, M.K.Y.; methodology, M.K.Y.; validation, M.K.Y. and A.A.; formal analysis, M.K.Y.; investigation, M.K.Y. and N.Y.A.; data curation, M.K.Y.; visualization, M.K.Y.; writing—original draft preparation, M.K.Y.; writing—review and editing, N.Y.A. and A.A.; supervision, N.Y.A. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by the German Academic Exchange Service (DAAD) under the programme Research Grants – Bi-nationally Supervised Doctoral Degrees/Cotutelle (Grant No. 57693451).

Data Availability Statement: The datasets generated and/or analyzed during the current study, together with the code used for simulation, surrogate-model training, and analysis, will be made publicly available in a permanent online repository upon acceptance of this manuscript. The supplementary Excel workbook and accompanying Supplementary Information describe the structure, harmonization, and coverage of the primary analysis dataset.

Acknowledgments: The authors acknowledge support from the Open Access Publication Fund of the University of Duisburg-Essen, the German Academic Exchange Service (DAAD), and the KNUST Engineering Education Project (KEEP).

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

 doi:10.20944/preprints202603.2296.v1

References

1. Singhania, R.R.; Patel, A.K.; Singh, A.; Haldar, D.; Soam, S.; Chen, C.W.; Tsai, M.L.; Dong, C.D. Consolidated bioprocessing of lignocellulosic biomass: Technological advances and challenges. *Bioresource Technology* **2022**, *354*, 127153. https://doi.org/10.1016/j.biortech.2022.127153.
2. Li, Z.; Waghmare, P.R.; Dijkhuizen, L.; Meng, X.; Liu, W. Research advances on the consolidated bioprocessing of lignocellulosic biomass. *Engineering Microbiology* **2024**, *4*, 100139. https://doi.org/10.1016/j.engmic.2024.100139.
3. Liu, Y.J.; Li, B.; Feng, Y.; Cui, Q. Consolidated bio-saccharification: Leading lignocellulose bioconversion into the real world. *Biotechnology Advances* **2020**, *40*, 107535. https://doi.org/10.1016/j.biotechadv.2020.107535.
4. Tsai, S.L.; Sun, Q.; Chen, W. Advances in consolidated bioprocessing using synthetic cellulosomes. *Current Opinion in Biotechnology* **2022**, *78*, 102840. https://doi.org/10.1016/j.copbio.2022.102840.
5. Sharma, J.; Kumar, V.; Prasad, R.; Gaur, N.A. Engineering of Saccharomyces cerevisiae as a consolidated bioprocessing host to produce cellulosic ethanol: Recent advancements and current challenges. *Biotechnology Advances* **2022**, *56*, 107925. https://doi.org/10.1016/j.biotechadv.2022.107925.
6. Minnaar, L.; den Haan, R. Engineering natural isolates of Saccharomyces cerevisiae for consolidated bioprocessing of cellulosic feedstocks. *Applied Microbiology and Biotechnology* **2023**, *107*, 7013–7028.
7. Periyasamy, S.; Beula Isabel, J.; Kavitha, S.; Karthik, V.; Mohamed, B.A.; Gizaw, D.G.; Sivashanmugam, P.; Aminabhavi, T.M. Recent advances in consolidated bioprocessing for conversion of lignocellulosic biomass into bioethanol – A review. *Chemical Engineering Journal* **2023**, *453*, 139783. https://doi.org/10.1016/j.cej.2022.139783.
8. Yeboah, M.K.; Asiedu, N.Y.; Dogbe, S.; Addo, A. Performance of Machine Learning Based-Modelling Approach in Consolidated Bioprocessing with Microbial Consortium for Bioethanol Production. *Industrial Biotechnology* **2024**, *20*, 77–97.
9. Maitra, S.; Singh, V. A consolidated bioprocess design to produce multiple high-value platform chemicals from lignocellulosic biomass and its technoeconomic feasibility. *Journal of Cleaner Production* **2022**, *377*, 134383.
10. Long, B.; Zhang, F.; Dai, S.Y.; Foston, M.; Tang, Y.J.; Yuan, J.S. Engineering strategies to optimize lignocellulosic biorefineries. *Nature Reviews Bioengineering* **2025**, *3*, 230–244.
11. Madhuvanthi, S.; Jayanthi, S.; Suresh, S.; Pugazhendhi, A. Optimization of consolidated bioprocessing by response surface methodology in the conversion of corn stover to bioethanol by thermophilic Geobacillus thermoglucosidasius. *Chemosphere* **2022**, *304*, 135242.
12. Yeboah, M.K.; Söffker, D. Consolidated Bioprocessing of Lignocellulosic Biomass: A Review of Experimental Advances and Modeling Approaches. *Bioresources and Bioproducts* **2026**, *2*. https://doi.org/10.3390/bioresourbioprod2010004.
13. Wen, Z.; Ledesma-Amaro, R.; Lu, M.; Jin, M.; Yang, S. Metabolic engineering of Clostridium cellulovorans to improve butanol production by consolidated bioprocessing. *ACS synthetic biology* **2020**, *9*, 304–315. https://doi.org/10.1021/acssynbio.9b00331.
14. Wen, Z.; Li, Q.; Liu, J.; Jin, M.; Yang, S. Consolidated bioprocessing for butanol production of cellulolytic Clostridia: development and optimization. *Microbial biotechnology* **2020**, *13*, 410–422.
15. Li, J.; Chen, B.; Gu, S.; Zhao, Z.; Liu, Q.; Sun, T.; Zhang, Y.; Wu, T.; Liu, D.; Sun, W.; et al. Coordination of consolidated bioprocessing technology and carbon dioxide fixation to produce malic acid directly from plant biomass in Myceliophthora thermophila. *Biotechnology for Biofuels* **2021**, *14*, 1–13.
16. Schlembach, I.; Hosseinpour Tehrani, H.; Blank, L.M.; Büchs, J.; Wierckx, N.; Regestein, L.; Rosenbaum, M.A. Consolidated bioprocessing of cellulose to itaconic acid by a co-culture of Trichoderma reesei and Ustilago maydis. *Biotechnology for biofuels* **2020**, *13*, 1–18.
17. Kumar, V.; Fox, B.G.; Takasuka, T.E. Consolidated bioprocessing of plant biomass to polyhydroxyalkanoate by co-culture of Streptomyces sp. SirexAA-E and Priestia megaterium. *Bioresource Technology* **2023**, *376*, 128934. https://doi.org/10.1016/j.biortech.2023.128934.
18. Fang, H.; Deng, Y.; Pan, Y.; Li, C.; Yu, L. Distributive and collaborative push-and-pull in an artificial microbial consortium for improved consolidated bioprocessing. *AIChE Journal* **2022**, *68*, e17844. https://doi.org/10.1002/aic.17844.
19. Helleckes, L.M.; Hemmerich, J.; Wiechert, W.; von Lieres, E.; Grünberger, A. Machine learning in bioprocess development: from promise to practice. *Trends in biotechnology* **2023**, *41*, 817–835.
20. Mondal, P.P.; Galodha, A.; Verma, V.K.; Singh, V.; Show, P.L.; Awasthi, M.K.; Lall, B.; Anees, S.; Pollmann, K.; Jain, R. Review on machine learning-based bioprocess optimization, monitoring, and control systems. *Bioresource technology* **2023**, *370*, 128523.

 doi:10.20944/preprints202603.2296.v1

21. Khanal, S.K.; Tarafdar, A.; You, S. Artificial intelligence and machine learning for smart bioprocesses, 2023.
22. Cheng, Y.; Bi, X.; Xu, Y.; Liu, Y.; Li, J.; Du, G.; Lv, X.; Liu, L. Artificial intelligence technologies in bioprocess: Opportunities and challenges. *Bioresource Technology* **2023**, *369*, 128451. https://doi.org/10.1016/j.biortech.2022.128451.
23. Agharafeie, R.; Oliveira, R.; Ramos, J.R.C.; Mendes, J.M. Application of hybrid neural models to bioprocesses: A systematic literature review. *Authorea Preprints* **2023**.