When More Sugar Makes Less Ethanol

Testing a Four-State Fermentation Model on Public Longan-Waste Data

Bob Zhao
The Webb Schools

Abstract

A model can trace known data beautifully and still fail on a new batch. I tested that problem with a public dataset of *Candida tropicalis* fermenting longan solid-waste hydrolysate. The data cover 12 concentration conditions, nine sampling times from 0 to 48 h, and four measured states: glucose, xylose, ethanol, and biomass. I built a four-state ordinary differential equation (ODE) model in which every condition shares the same parameter set and differs only through its measured starting concentrations. I compared three versions of the model: production only, ethanol reassimilation, and a version that also lowers apparent ethanol yield under strong matrix stress. Each version used three possible stress-curve shapes, producing nine candidates in total. Every candidate had to predict an entire concentration condition that had been left out of fitting. The winner was *M_2* with *h_I=2*. Its internal model-selection score was 13.06% mean held-out nRMSE, with statewise values of 8.2% for glucose, 7.9% for xylose, 17.4% for ethanol, and 18.7% for biomass. A 200-draw parametric bootstrap was used to examine uncertainty after the model was selected. The model captured broad patterns across much of the concentration range, but it struggled near factors 3 and 5, where the experiments move from productive fermentation to strong suppression. In other words, the model learned the main pattern, but it definitely did not learn everything. The results support a useful computational description of this dataset, not a universal law of fermentation or proof of one exact biological mechanism.

1 Background

1.1 Fermentation as a Dynamical Systems Problem

Fermentation may look quiet from outside the flask, but the chemistry is constantly moving. Sugars disappear, cells grow, ethanol builds up, and sometimes that ethanol later decreases. A time-course experiment is therefore not just a pile of measurements. It is a partial movie of a connected system. My goal was to describe that movie with a small set of equations and then see whether the same equations could handle a batch they had never seen.

The dataset comes from Feng et al., who fermented longan solid-waste hydrolysate with *Candida tropicalis* TISTR 5306 and published the time-course tables as supplementary data [1]. Longan waste is the solid material left after juice processing. It contains glucose and xylose, but it also contains phenolic compounds and other parts of the hydrolysate matrix. That makes it interesting for modeling: adding more feedstock adds more sugar, but it can also make the environment harder for the yeast.

The data quickly showed that "more sugar equals more ethanol" was too simple. Glucose usually disappeared first, ethanol rose and sometimes fell again, and the most concentrated conditions barely fermented at all. The yeast did not seem impressed by all the extra sugar. I therefore needed three basic ideas in the model: sugar uptake that levels off, a way to represent stress in concentrated hydrolysate, and a possible switch from producing ethanol to consuming it.

Classical kinetic ideas gave me a starting point. Monod-type saturation describes how uptake or growth approaches a maximum as more substrate becomes available [2]. Yeast can also reorganize its carbon use after glucose runs out [3]. Finally, high sugar concentrations can create osmotic stress and reduce yeast performance [4]. These studies make the model choices reasonable, but they do not prove that the same mechanisms caused the longan results.

1.2 The Difference Between Fit and Transfer

A fitted curve can be a bit of a show-off: looking good on familiar data is much easier than predicting a missing batch. A low training error might mean the model found a reusable rule, or it might mean the model became very good at copying the conditions it already knew. I cared more about the first possibility.

Time-course data make this tricky. A random split of individual time points would place measurements from the same concentration trajectory in both training and testing. The "test" points would still sit beside measurements the model had already seen. Cross-validation should match the grouped prediction problem [5], so I left out one complete concentration trajectory at a time, fitted the other 11, and predicted all four variables at the eight later times from the missing condition's starting state.

This is a harder test because the curves do not simply become taller as concentration increases. Factors 1.0 through 3.0 are productive, while factors 5.0 and 7.0 are strongly suppressed. If one of the transition conditions is missing, the model must figure out what happens between two very different regimes. That is exactly where I wanted to see whether it could keep its balance.

1.3 Longan-Waste Data as a Public Test Case

The source study performed 12 batch fermentations at concentration factors 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.8, 2.0, 3.0, 5.0, and 7.0 [1]. Initial glucose plus xylose ranged from about 30.6 to 213.0 g/L, while the glucose-to-xylose ratio stayed close to 2.1-2.2:1. Samples were collected every 6 h for 48 h. The tables report means and standard errors from five replicates for glucose, xylose, ethanol, dried biomass, and four phenolic compounds.

The wide concentration range makes the dataset useful, but there is a catch. Only summary means and standard errors are public, not the five individual trajectories. Also, nearly every major matrix component rises with concentration factor. Initial total sugar has correlations above 0.9998 with each

measured phenolic and with their sum. Sugar and phenolics are riding together so closely that this experiment cannot tell which passenger is driving the inhibition.

For that reason, I call the model term **effective aggregate matrix stress**. It summarizes a clear concentration-related pattern, but it does not prove that sugar alone caused the stress. The equation is allowed to describe what happened without pretending the experiment measured more than it did.

1.4 Research Question and Contributions

The central question is:

> *Can one compact four-state ODE, using one parameter set for all conditions, predict glucose, xylose, ethanol, and biomass when a whole concentration condition is left out of fitting?*

To answer the question, I reconstructed the public tables, built a four-state ODE, compared nine candidates, tested them on whole missing conditions, and examined uncertainty with a bootstrap. Throughout the analysis, I kept fitted and held-out performance separate because they answer different questions.

I was not trying to discover one permanent kinetic constant for *C. tropicalis*. This dataset cannot support that claim. I wanted a useful map of where a compact model works, where it gets lost, and what experiment would help next.

2 Model Concept

2.1 Four Observable States

The model follows four concentrations that appear in every table:

- $G(t)$: glucose concentration in g/L,
- $Z(t)$: xylose concentration in g/L,
- $E(t)$: ethanol concentration in g/L, and
- $X(t)$: dried biomass concentration in g/L.

I kept the phenolic measurements for descriptive analysis, but I did not add them as separate dynamic drivers. Their starting values track the sugar series almost perfectly, so separate inhibition coefficients would look more detailed without actually being more informative.

Each condition starts from its own reported $G(0)$, $Z(0)$, $E(0)$, and $X(0)$. After that, every condition follows the same ODE and the same estimated parameters. The concentration-factor label never enters the equation as a correction. A missing condition must therefore be predicted from its measured starting state alone.

2.2 Substrate Uptake and Aggregate Matrix Stress

Glucose and xylose have separate uptake fluxes that level off as substrate becomes abundant. A shared decreasing function of total sugar, $G+Z$, represents the loss of performance in concentrated hydrolysate. At low total sugar the multiplier is close to one. At high total sugar it slows both uptake rates.

I tested three fixed Hill exponents: $h\_I=2$, 4, and 8. The lower value creates a smooth transition, while the higher value behaves more like a switch. I treated these exponents as different curve shapes, not biological measurements. If $h\_I=2$ wins, it means the smoother option predicted this panel better.

2.3 Ethanol Formation and Reassimilation

The model forms ethanol from the combined glucose and xylose uptake flux. Its apparent ethanol yield can decrease as matrix stress increases. This lets a concentrated batch consume some sugar without automatically producing ethanol in the same proportion as a moderate batch.

The second idea is reassimilation. Once glucose becomes very small, an ethanol-uptake term can turn on and contribute to biomass. Yeast carbon-use transitions after glucose depletion have been studied before [3], so this is biologically reasonable. Still, the samples are 6 h apart. They cannot reveal the exact instant when a fast switch occurs.

For that reason, I fixed *K_switch=0.1* g/L instead of estimating it. It is a numerical gate that opens near glucose depletion, not a directly measured physiological threshold.

2.4 A Flux-Linked, Not Closed-Balance, Model

The four equations share the same substrate and ethanol fluxes. Sugar uptake feeds modeled ethanol and biomass formation, while ethanol uptake removes ethanol and can add biomass. This keeps product formation connected to material uptake.

The model is not a complete carbon balance. It leaves out carbon dioxide, maintenance, respiration, cell death, storage, and unmeasured metabolites. Because those pathways are missing, I call the model **flux-linked** rather than fully mass-balanced. The equations share material flows, but they do not follow every carbon atom through the flask.

2.5 Nested Architectural Questions

Three models ask progressively richer questions:

1. **Production-only model (M_0).** Can saturating, stress-limited sugar uptake and constant apparent yields describe all four trajectories without ethanol loss?
2. **Reassimilation model (M_1).** Does adding ethanol uptake and ethanol-supported biomass improve transfer, especially after the ethanol peak?
3. **Stress-switch model (M_2).** Does allowing the apparent ethanol yield to decline with aggregate stress provide further predictive value?

Each model adds one layer of behavior. The extra flexibility only counts as useful if it improves prediction. I therefore chose models first by held-out performance and used information criteria as a second opinion. Figure 1 summarizes the measured fermentation landscape, and Figure 2 shows how the model links the four states.

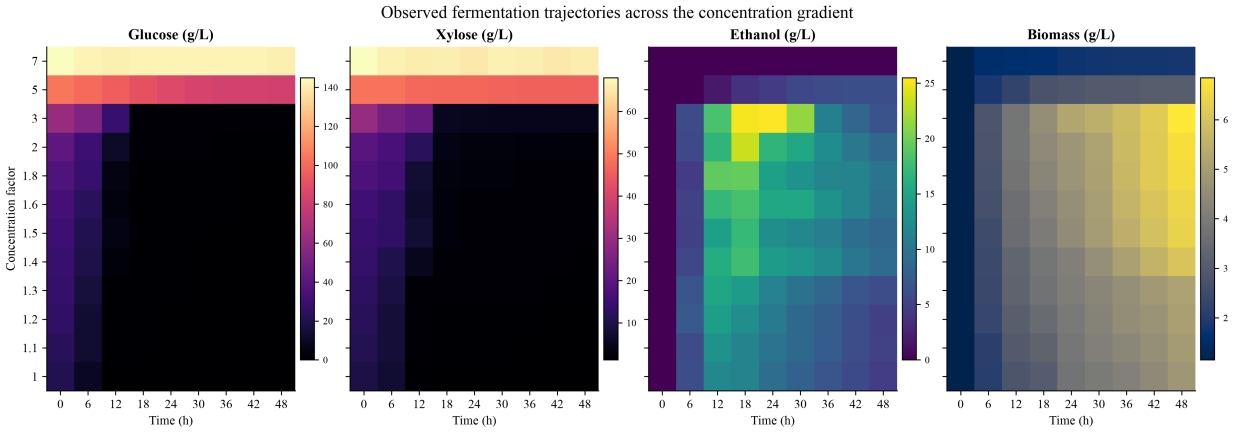

Figure 1: Observed fermentation trajectories across the concentration gradient. Each cell is the reported quintuplicate mean at one concentration factor and sampling time. Productive ethanol formation is concentrated in the intermediate loading range, whereas factors 5 and 7 retain substrate and show strongly reduced biomass and ethanol formation.

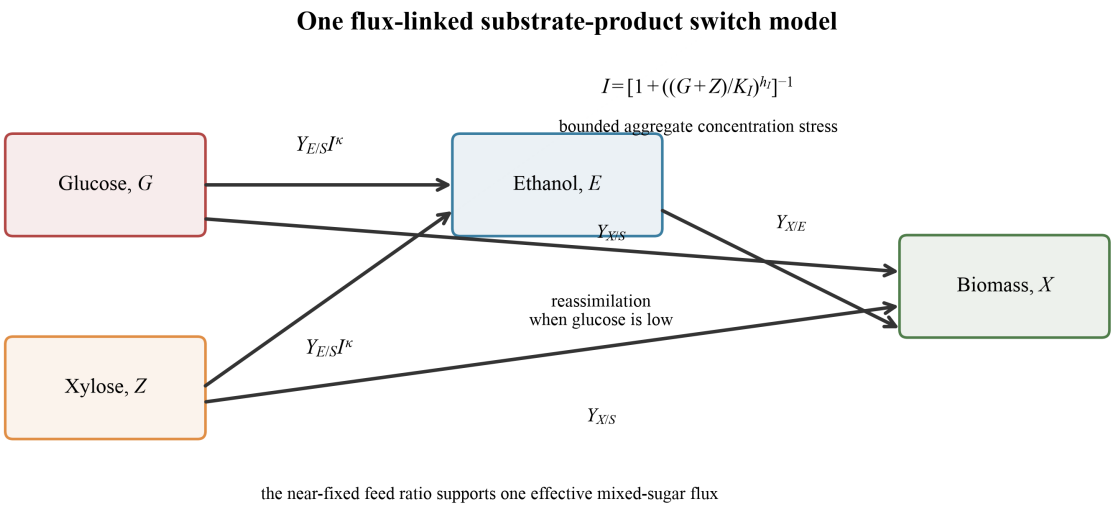

Figure 2: Flux-linked four-state model. Glucose and xylose uptake feed apparent ethanol and biomass formation; ethanol can support biomass only after the fixed near-zero-glucose gate opens. The bounded multiplier I represents aggregate matrix stress. Carbon dioxide, oxygen, maintenance, and unmeasured products are omitted, so the diagram is not a closed elemental balance.

3 Methodology

3.1 Data Source and Experimental Panel

I used the supplementary kinetics tables from Feng et al. [1]. I did not perform new wet-laboratory experiments. The original study used *C. tropicalis* TISTR 5306 and longan solid-waste hydrolysate at 12 concentration factors, with measurements at 0, 6, 12, 18, 24, 30, 36, 42, and 48 h. The supplement reports standard errors for the four modeled variables and also includes four phenolic compounds.

The PDF was readable to a person and much less friendly to a computer. Its tables do not have ruling lines, so ordinary table-detection tools miss the structure. I wrote an extraction script that uses the stable column positions, assigns each number to the nearest expected column, and checks every time sequence. The final data frame has 108 rows and 18 columns. I verified the 12×9 shape, nonnegative values, complete time grids, and several values from different pages, including 11.9 g/L ethanol at factor 1.0 and 12 h, 25.5 g/L at factor 3.0 and 24 h, and no ethanol increase at factor 7.0.

Table 1 summarizes the panel. One concentration condition is the experimental unit. Each published point summarizes five replicates, but the five individual trajectories are not available. I therefore treated the published condition-level means as the available observations and did not reconstruct synthetic replicates from the reported standard errors.

Table 1. Public longan-waste fermentation panel.

Feature	Value used in this analysis
Organism	*Candida tropicalis* TISTR 5306
Feedstock	Longan solid-waste hydrolysate
Concentration factors	1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.8, 2.0, 3.0, 5.0, 7.0
Initial total sugar range	Approximately 30.6-213.0 g/L
Initial glucose:xylose ratio	Approximately 2.1-2.2:1
Sampling times	0 to 48 h in 6 h increments
Number of trajectories	12
Time points per trajectory	9
Modeled states	Glucose, xylose, ethanol, biomass
Reported uncertainty	Mean and standard error from quintuplicate experiments
Noninitial fitted observations	*12×8×4=384*
Additional measured variables	Gallic acid, ellagic acid, pyrogallol, catechol
Primary source	Feng et al. [1]

3.2 State Equations

For concentration condition c, let

$$\mathbf{y}_c(t) = [G_c(t), Z_c(t), E_c(t), X_c(t)]^{\mathrm{T}}$$

The measured state at $t=0$ starts each simulation, and the ODE supplies every later value. Rate expressions use nonnegative states, and tiny negative concentrations produced by the numerical solver are clipped to zero.

The effective aggregate matrix-stress function is

$$I(G, Z; K_I, h_I) = \frac{1}{1 + \left(\frac{G+Z}{K_I}\right)^{h_I}} \tag{1}$$

where $K\_I>0$ is an estimated concentration scale and h_I is fixed at 2, 4, or 8 for each candidate. Equation (1) uses total sugar because it is measured and orders the concentration series. It does not prove that sugar alone caused the inhibition.

The specific substrate-uptake fluxes are

$$q_G = q_{G,\max} \frac{G}{K_G + G} I(G, Z) \tag{2}$$

and

$$q_Z = q_{Z,\max}\frac{Z}{K_Z + Z}R(G)I(G,Z) \tag{3}$$

An exploratory glucose-repression multiplier was written as

$$R(G) = \rho + \frac{1-\rho}{1+(G/K_R)^{h_R}} \tag{4}$$

Preliminary optimization pushed ρ to 1, so I fixed $\rho=1$ and made $R(G)$ equal to 1. In plain language, the fitted model contains **no glucose-repression effect**. Equation (4) is included to show what was tested and removed. The organism may still prefer glucose; this particular dataset and model simply did not support a separate repression coefficient.

Ethanol reassimilation is gated by

$$S_G(G) = \frac{1}{1+(G/K_{\text{switch}})^{h_{\text{switch}}}} \tag{5}$$

and the specific ethanol-uptake flux is

$$q_E = k_E E S_G(G) \tag{6}$$

The apparent ethanol yield is

$$Y_{E/S,\text{app}} = Y_{E/S} I(G,Z)^{\kappa} \tag{7}$$

Finally, the four state equations are

$$\frac{dG}{dt} = -q_G X \tag{8}$$

$$\frac{dZ}{dt} = -q_Z X \tag{9}$$

$$\frac{dE}{dt} = [Y_{E/S,\text{app}}(q_G + q_Z) - q_E]X \tag{10}$$

and

$$\frac{dX}{dt} = [Y_{X/S}(q_G + q_Z) + Y_{X/E}q_E]X \tag{11}$$

Equations (8)-(11) share uptake fluxes, but they still omit carbon dioxide, maintenance, respiration, death, and other metabolites. The apparent yields connect substrate uptake to modeled ethanol and biomass changes; they are not a complete chemical accounting.

3.3 Fixed Quantities and Estimated Parameters

I fixed several constants before comparing models because the 6 h sampling grid could not identify everything at once. The half-saturation constants are $K\_G=K\_Z=1.0$ g/L. The unused repression expression keeps $K\_R=2.0$ g/L and $h\_R=4$, although $\rho=1$ makes them irrelevant to prediction. The ethanol gate uses $K\_switch=0.1$ g/L and $h\_switch=4$. Since glucose can move from a positive measurement to zero between two samples, the exact gate threshold is not identified by these data.

The active parameters depend on the candidate architecture. Table 2 records the distinction.

Table 2. Candidate architectures and parameters. A dash indicates that the mechanism is disabled. Fixed values are assumptions used to stabilize the global fit, not experimentally estimated constants.

Quantity	Interpretation	*M_0* production only	*M_1* reassimilation	*M_2* stress-switch
q_G,max	Maximum specific glucose-uptake coefficient	Estimated	Estimated	Estimated
q_Z,max	Maximum specific xylose-uptake coefficient	Estimated	Estimated	Estimated
K_I	Aggregate matrix-stress concentration scale	Estimated	Estimated	Estimated
Y_E/S	Low-stress apparent ethanol yield	Estimated	Estimated	Estimated
Y_X/S	Apparent biomass yield from sugar	Estimated	Estimated	Estimated
k_E	Ethanol reassimilation coefficient	-	Estimated	Estimated
Y_X/E	Apparent biomass yield from reassimilated ethanol	-	Estimated	Estimated
κ	Stress dependence of apparent ethanol yield	-	-	Estimated
h_I	Shape of aggregate-stress curve	Fixed candidate: 2, 4, or 8	Fixed candidate: 2, 4, or 8	Fixed candidate: 2, 4, or 8
K_G,K_Z	Substrate half-saturation constants	Fixed at 1.0 g/L	Fixed at 1.0 g/L	Fixed at 1.0 g/L
ρ	Residual xylose-uptake factor under glucose	Fixed at 1; no repression	Fixed at 1; no repression	Fixed at 1; no repression
K_switch	Numerical glucose gate for ethanol use	-	Fixed at 0.1 g/L	Fixed at 0.1 g/L
h_switch	Sharpness of numerical ethanol gate	-	Fixed at 4	Fixed at 4

The number of active kinetic parameters is five in *M_0*, seven in *M_1*, and eight in *M_2*. Four additional state-specific discrepancy scales are estimated for the likelihood-based comparison described below.

3.4 Numerical Integration and Objective Function

I integrated the ODE with SciPy's solve_ivp implementation of the DOP853 solver. Relative and absolute tolerances were *10^(-5)* and *10^(-7)*. NumPy handled the arrays and data transformations. I estimated parameters with bounded nonlinear least squares. Rate parameters received log-uniform random starting values, while bounded yields and shape parameters received uniform starts inside their allowed ranges.

The published standard errors are much smaller than the mismatch between one global ODE and all 12 trajectories. If I weighted only by those errors, a few tiny or zero values would dominate the fit. I therefore added a state-specific discrepancy floor. For state j, condition c, and time t, the residual scale is

$$\sigma_{ctj} = \sqrt{\mathrm{SE}_{ctj}^2 + \tau_j^2} \tag{12}$$

where $\tau_j \geq 0$ is a state-specific discrepancy floor. The standardized residual is

$$r_{ctj}(\theta) = \frac{\hat{y}_{ctj}(\theta) - y_{ctj}}{\sigma_{ctj}} \tag{13}$$

I alternated between updating the kinetic parameters and updating the discrepancy scales. After each parameter fit, τ_j was updated from residual variance beyond the reported measurement variance, with a small lower bound. The loop stopped when the values stabilized or reached the iteration limit. This gives model mismatch some room, but it is not a complete probability model for correlated fermentation errors.

Time zero is not included in the residual vector because it is supplied exactly as the initial condition. Each full-data objective therefore contains $12 \times 8 \times 4 = 384$ standardized noninitial observations. Optimization used a linear least-squares loss, Jacobian scaling, bounded parameters, deterministic starts, and reproducible random starts based on seed 20260830. Full-data candidates used six multistarts. Each blocked-validation fit used three fold-local multistarts and was never initialized from the corresponding full-data optimum.

3.5 Complete Candidate Grid

The three architectures were each tested with $h_I = 2$, 4, and 8, giving nine candidates. I tested every exponent for every architecture so no model received an extra tuning advantage. Otherwise, more chances to improve could easily be mistaken for a better equation.

Every candidate received one full-data fit and 12 held-out-condition fits. The winner minimized the equal-state mean leave-one-concentration-out normalized root-mean-square error. Because I used the same 12 folds to choose the model and report its score, the result is an **internal model-selection score**. It may be optimistic compared with performance on a completely new experiment. It is not external validation.

3.6 Leave-One-Concentration-Out Prediction

In fold c, I withheld all nine time points from condition c. The other 11 conditions estimated the parameters from fold-local starting values. The held-out condition's simulation then began at its measured $G_c(0)$, $Z_c(0)$, $E_c(0)$, and $X_c(0)$ and predicted the other eight times. The ODE did not get to peek at the missing condition's label.

Prediction error is summarized by state using root-mean-square error,

$$\mathrm{RMSE}_j = \sqrt{\frac{1}{N_j}\sum_i (\hat{y}_{ij} - y_{ij})^2} \tag{14}$$

mean absolute error,

$$\mathrm{MAE}_j = \frac{1}{N_j}\sum_i |\hat{y}_{ij} - y_{ij}| \tag{15}$$

and normalized RMSE,

$$\mathrm{nRMSE}_j = \frac{\mathrm{RMSE}_j}{\max(y_j) - \min(y_j)} \tag{16}$$

The normalization range is computed globally for each state, so a nearly flat held-out condition does not produce an undefined or arbitrarily large condition-specific denominator. *R^2* is also reported, but it is interpreted cautiously for flat trajectories. Candidate selection uses the arithmetic mean of the four statewise nRMSE values, giving glucose, xylose, ethanol, and biomass equal influence after scaling.

Blocked validation is better suited to these grouped trajectories than a random time-point split [5]. It still has limits: there are only 12 folds, nearby concentration factors are related, and the selection procedure can favor a model that happens to fit this particular grid.

3.7 Approximate Information-Criterion Comparison

I also compared the full-data fits using corrected Akaike information criterion (AICc). For this calculation, I treated the standardized residuals as conditionally Gaussian with the scale from Equation (12). The resulting negative log-likelihood is used to calculate

$$\mathrm{AICc} = 2k - 2\log L + \frac{2k(k+1)}{n-k-1} \tag{17}$$

where *n=384* and *k* counts active kinetic parameters plus the four discrepancy scales. The small-sample correction follows Hurvich and Tsai [6].

This AICc is approximate for two reasons. Measurements from one fermentation are related over time, while Equation (17) treats the residuals as conditionally independent. Also, the discrepancy scales come from an iterative procedure instead of a full hierarchical likelihood. I therefore used AICc as a second opinion about fit and complexity, not as the main selection rule. If AICc and held-out prediction disagree, that is not a calculation error. They are asking different questions.

3.8 Conditional Parametric Bootstrap

After selecting the candidate, I ran a 200-draw parametric bootstrap. For each draw, I added independent Gaussian perturbations to the selected model's full-data predictions using the combined scales in Equation (12). Concentrations were constrained to remain nonnegative, and the measured initial states were restored. The selected architecture and *h_I* remained fixed, and each synthetic dataset was refitted from the selected full-data estimate. Parameter distributions and dense-grid peak metrics were then summarized with percentile intervals.

The bootstrap is conditional on the selected setup. It does not repeat extraction, resample the unavailable raw replicates, choose a new architecture, choose a new Hill exponent, or account for every process missing from the equations. All 200 fits succeeded. That is reassuring, but it is not a magic wand: the intervals describe uncertainty inside the chosen model, not uncertainty across every possible explanation.

I also screened practical identifiability with singular values of the scaled least-squares Jacobian, its condition number, and an approximate parameter-correlation matrix from the pseudoinverse of *J^TJ*. These local checks can reveal parameter combinations that are hard to estimate, but they cannot prove that the parameters are unique. A formal profile-likelihood analysis would be needed before treating the fitted coefficients as uniquely determined kinetic constants [7].

3.9 Derived Harvest Metrics and Reproducibility

For each observed concentration condition, peak ethanol, time of the observed peak, net productivity to the peak, and net ethanol yield over sugar consumed are calculated directly from the reported grid. The fitted ODE is also integrated on a 0.2 h grid from 0 to 48 h to obtain model-derived peak quantities. These dense-grid values refine the numerical location of a peak within the model; they do not add experimental temporal resolution. In particular, a model-derived peak between two 6 h samples remains an interpolation of the assumed equations.

All analysis is scripted. The source supplement, extracted data, model code, candidate scores, held-out predictions, bootstrap summaries, and figures use a fixed random seed. Scientific computing uses SciPy and NumPy, and the figures are generated with Matplotlib. Another person can rerun the procedure, but code cannot recover raw replicates that were never published or turn internal cross-validation into an external experiment.

4 Results and Validation

4.1 Observed Concentration Regimes

The raw data do not show a simple “more starting sugar, better fermentation” relationship. From factors 1.0 through 2.0, the ethanol peak generally rises and shifts later, from 12 h in the most dilute conditions to 18 h in several intermediate ones. Factor 3.0 starts with 91.8 g/L total sugar and reaches the largest measured ethanol concentration, 25.5 g/L at 24 h. Its net peak yield is lower because a smaller share of the sugar consumed appears as ethanol by the peak.

Then the pattern changes sharply. At factor 5.0, ethanol peaks at only 6.1 g/L and not until 42 h. Final biomass drops to 3.08 g/L, less than half the 6.85 g/L at factor 3.0. At factor 7.0, ethanol never rises above its 0.02 g/L starting value, final biomass is only 1.82 g/L, and plenty of sugar remains at 48 h. More feedstock produced much less ethanol-not exactly the bargain I expected.

Table 3 reports representative quantities calculated directly from the public observations rather than from the model.

Table 3. Representative observed harvest metrics. Net productivity is the increase in ethanol divided by observed peak time. Net yield is ethanol increase divided by sugar consumed by the observed peak.

Concentration factor	Initial total sugar (g/L)	Peak ethanol (g/L)	Peak time (h)	Net productivity (g/L/h)	Net peak yield (g/g)	Final biomass (g/L)
1.0	30.58	11.9	12	0.985	0.408	4.82
1.4	42.3	17.7	18	0.982	0.453	6.03
2.0	60.3	23.5	18	1.303	0.426	6.66
3.0	91.8	25.5	24	1.061	0.300	6.85
5.0	152.5	6.1	42	0.145	0.249	3.08
7.0	213.0	0.02	-	-	Not defined	1.82

The phenolic measurements rise along the same concentration series. Initial total sugar has correlations of about 0.9999 with each initial phenolic and with their sum. Depending on the equation, a regression could assign the effect to either group. The safest conclusion is that the **concentrated hydrolysate matrix** is associated with inhibition; the data do not identify one causal compound.

4.2 Full-Data Reconstruction

The prediction-first rule selected *M_2* (stress-dependent apparent yield with ethanol reassimilation) with *h_I=2*. Its estimated parameter vector is *q_G,max*=2.238 g sugar g^*(-1)_X* h^*(-1)*; *q_Z,max*=0.5575 g sugar g^*(-1)_X* h^*(-1)*; *K_I*=62.09 g/L; *Y_E/S*=0.4657 g ethanol/g sugar; *Y_X/S*=0.04941 g biomass/g

sugar; k_E=0.006407 L g^(-1)_X h^(-1); $Y_{X/E}$=0.1608 g biomass/g ethanol; and κ=0.7298. The fitted discrepancy scales are $\tau_G=5.05$, $\tau_Z=2.94$, $\tau_E=2.89$, and $\tau_X=0.645$ g/L. Figure 3 compares the observed data with representative fitted curves.

The fitted curves capture the main substrate patterns. In-sample R^2 is 0.987 for glucose and 0.979 for xylose, but only 0.744 for ethanol and 0.781 for biomass. Full-data nRMSE is 3.6% for glucose, 4.5% for xylose, 11.4% for ethanol, and 12.1% for biomass. The main misses are the shape of ethanol and biomass curves, leftover xylose in dilute batches, and the timing and biomass level near factors 3 and 5. These are fitting results after every condition contributed information, so they should not be mistaken for new-condition prediction.

The discrepancy floors are much larger than many reported analytical standard errors. Most of the remaining error therefore comes from model mismatch and differences between conditions, not just laboratory measurement noise. This is why using only the tiny published errors would make the fit look far more certain than it really is.

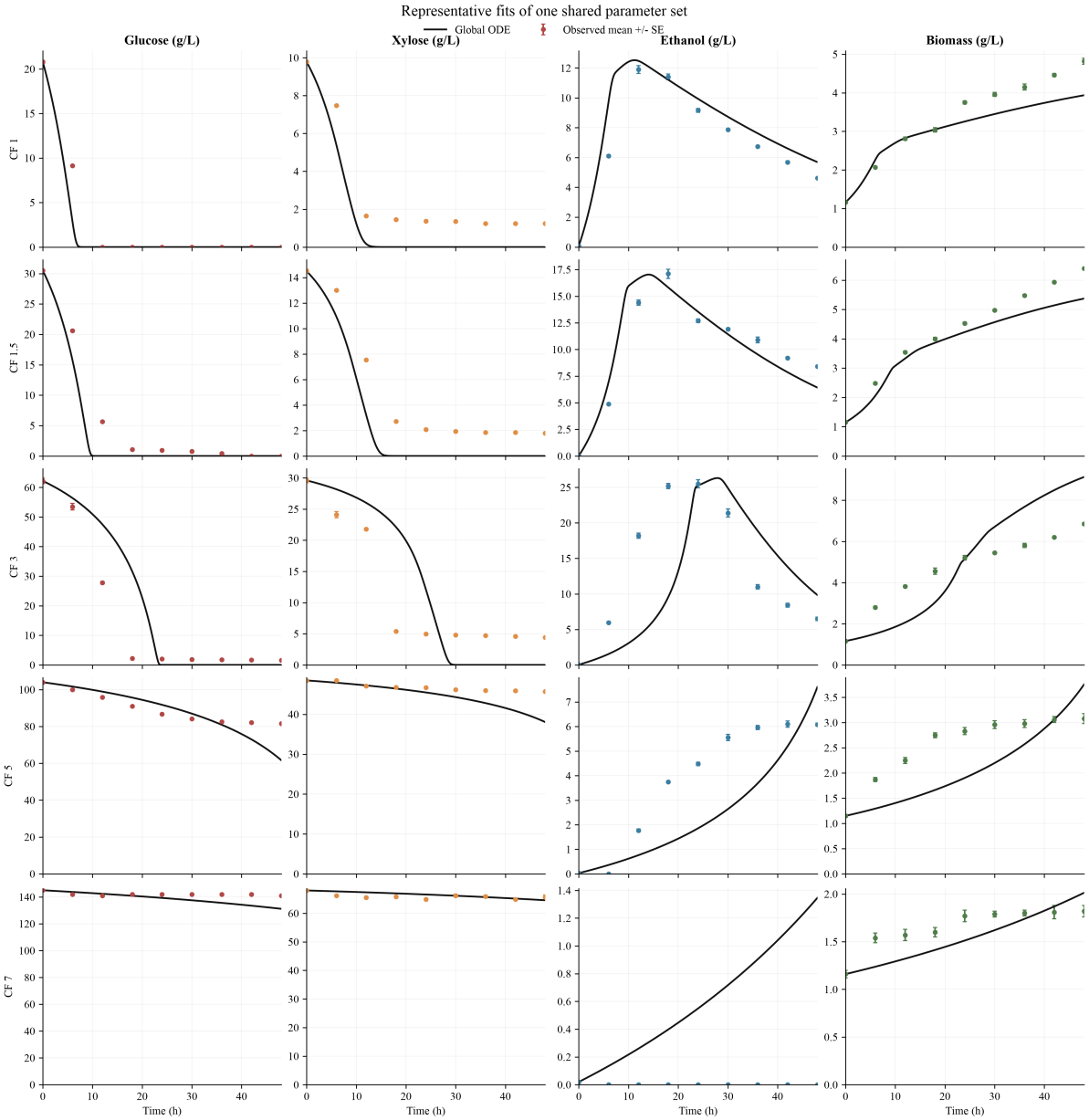

Figure 3: Representative in-sample reconstructions from one parameter vector shared by all conditions. Points are reported means with standard errors; black curves are the selected full-data ODE fit. The rows span dilute, intermediate, transition, and strongly inhibited conditions. These curves diagnose fit after calibration and are not held-out predictions or uncertainty bands.

4.3 Prediction-First Candidate Comparison

Across the nine candidates, mean blocked nRMSE ranged from 13.06% to 24.84%. All three architectures preferred the smooth *h_I=2* curve under the joint-state score. The best versions of *M_0*, *M_1*, and *M_2* scored 16.00%, 14.48%, and 13.06%, respectively. Nine candidates were tested, and the held-out conditions chose *M_2,h_I=2*.

The selected model's statewise held-out nRMSE values are 8.2% for glucose, 7.9% for xylose, 17.4% for ethanol, and 18.7% for biomass. The matching *R^2* values are 0.928, 0.935, 0.403, and 0.477. The added terms did not help every state equally. Moving from *M_0* to *M_1* improved glucose, xylose, and biomass but made ethanol slightly worse, from 23.6% to 25.1% nRMSE. Adding stress-dependent yield in *M_2* lowered ethanol error to 17.4%, while glucose and xylose became slightly worse and biomass barely changed. The winner is therefore the best tested compromise under an equal-state score, not necessarily the best model under every possible metric.

The AICc ranking told a different story. It preferred *M_2,h_I=8* (1509.9), followed by *M_1,h_I=8* (1516.6), while the prediction-selected *M_2,h_I=2* fit had AICc 1751.7. AICc and blocked prediction disagreed. That was awkward, but scientifically useful. The sharp curve fit the full panel more closely, while the smooth curve transferred better when a complete condition was missing. Figure 4 shows the held-out predictions, and Figure 5 puts the two rankings side by side.

Table 4. Final candidate-grid summary. All nine candidates are retained so that tuning opportunities and the fit-prediction conflict remain visible.

Architecture	*h_I*	Active kinetic parameters	Mean LOCO nRMSE (fraction)	Glucose nRMSE	Xylose nRMSE	Ethanol nRMSE	Biomass nRMSE	Approximate AICc
M_0	2	5	0.1600	0.0784	0.0750	0.2363	0.2502	1894.7
M_0	4	5	0.1801	0.0843	0.0960	0.2459	0.2943	1718.8
M_0	8	5	0.2484	0.1252	0.1464	0.3150	0.4070	1700.0
M_1	2	7	0.1448	0.0749	0.0662	0.2510	0.1870	1797.2
M_1	4	7	0.1662	0.0752	0.0866	0.2458	0.2574	1563.8
M_1	8	7	0.2409	0.1229	0.1410	0.2843	0.4152	1516.6
M_2	2	8	0.1306	0.0821	0.0790	0.1739	0.1875	1751.7
M_2	4	8	0.1495	0.0841	0.0954	0.1567	0.2617	1536.6
M_2	8	8	0.2157	0.1237	0.1427	0.1977	0.3987	1509.9

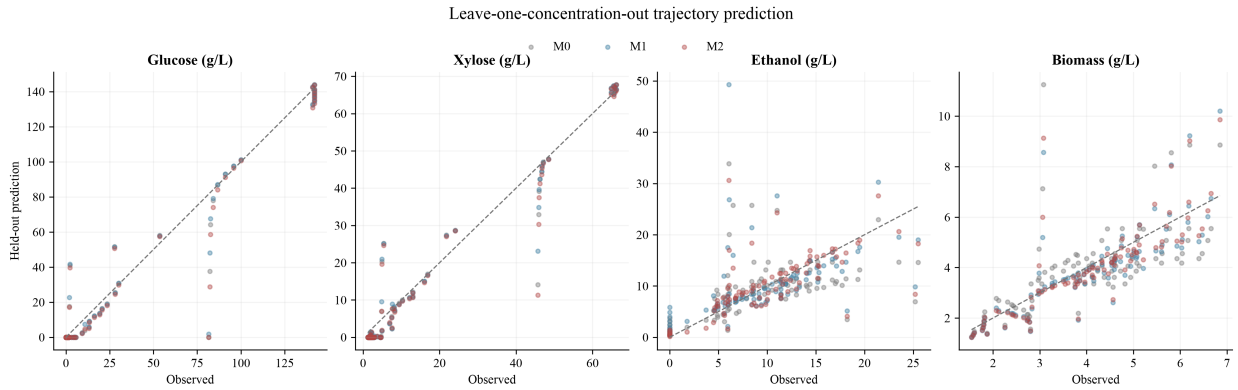

Figure 4: Leave-one-concentration-out predictions versus observations. Every plotted prediction comes from a fit that excluded the complete corresponding concentration trajectory. The dashed line is identity. The tuned h_I=2 version of each architecture is shown; high-leverage departures in ethanol and biomass expose the weaker transfer of those states.

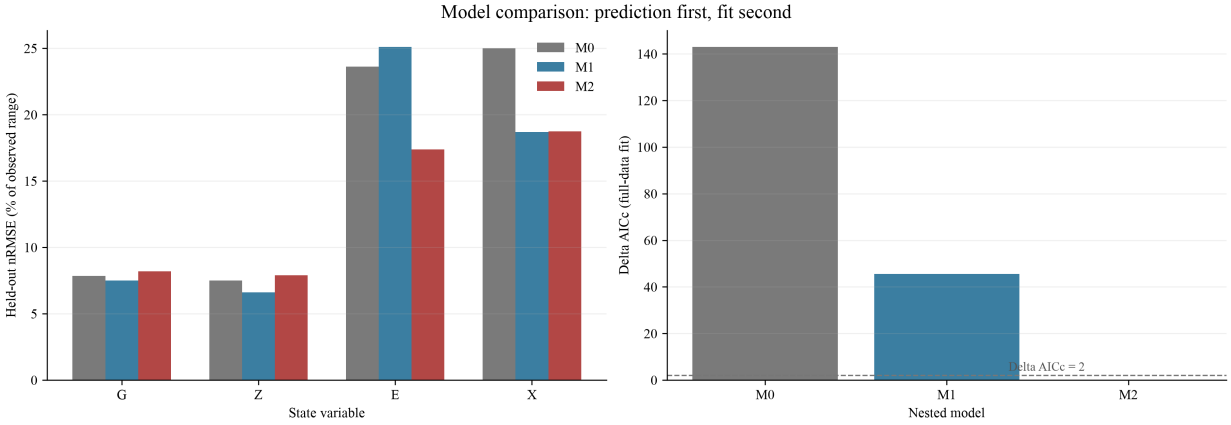

Figure 5: Prediction-first architecture comparison. Left: statewise blocked nRMSE for each architecture after its own Hill-exponent tuning. Right: approximate full-data delta AICc. All architectures selected h_I=2 for blocked joint-state prediction, while sharper curves fit the full panel more closely. AICc is secondary because discrepancy scales are iterative and residual independence is uncertain.

4.4 Where Condition Transfer Fails

One average score can hide where a model really trips. The largest held-out errors occur at factors 3 and 5, as Figure 6 shows. These conditions sit on opposite sides of the change from productive fermentation to severe suppression. Without factor 3, the model has to guess the highest ethanol peak from lower factors and the very different factor-5 curve. Without factor 5, it must locate the start of suppression between factors 3 and 7. The smooth stress curve does not always place that boundary correctly.

At factor 3, nRMSE was 11.6% for glucose, 14.4% for xylose, 41.9% for ethanol, and 37.9% for biomass. At factor 5, the values were 25.2%, 21.3%, 37.5%, and 46.0%. These were not small vertical shifts. In the factor-5 fold, the model predicted complete glucose depletion and 30.6 g/L final ethanol, while the data still had 81.6 g/L glucose and only 6.08 g/L ethanol. It also predicted 9.87 g/L final biomass at factor 3 instead of the observed 6.85 g/L. Factor 5 is where the fitted equations disagree most clearly with the held-out data.

This is also why I do not treat the winning stress exponent as a biological discovery. It may win because its smoother curve acts like regularization across a poorly sampled transition. There are no direct measurements between factors 3 and 5, and the components of the hydrolysate were not varied independently.

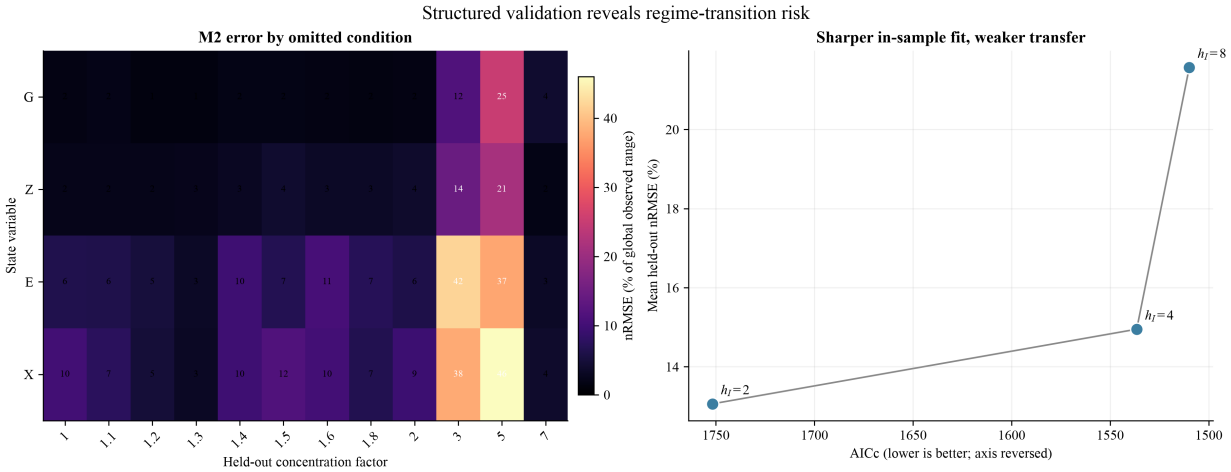

Figure 6: Structured validation reveals regime-transition risk. The left heatmap reports selected-model nRMSE for each omitted concentration and state; factors 3 and 5 dominate the practical failures. The right panel shows that, within M2, lower full-data

AICc accompanies worse blocked transfer as h_I sharpens. The same folds were used for selection and scoring, so these are internal, selection-biased model-comparison results.

4.5 Harvest Trade-Offs

Figure 7 shows a trade-off between titer, time, and yield. Factor 2.0 has higher observed net productivity than factor 3.0 because it reaches 23.5 g/L ethanol by 18 h, while factor 3.0 needs 24 h to reach 25.5 g/L. Factor 1.4 has the best representative peak yield in Table 3, and factor 3.0 has the largest titer. Factor 5.0 has much more starting sugar and performs poorly by every measure.

On the dense model grid, factor 2 peaks at 20.94 g/L and 17.4 h, while factor 3 peaks at 26.32 g/L and 27.8 h. Factors 5 and 7 reach their simulated maximum at the 48 h boundary, so those times are censored rather than true interior optima. For factor 3, the conditional bootstrap interval is 24.09-28.81 g/L for peak ethanol and 26.2-29.4 h for peak time. The factor-5 and factor-7 peak times stay at 48 h. A narrow interval around a poor prediction shows parameter stability within the chosen model, not accuracy of the model form. This is why the observed points remain visible in Figure 7.

I did not optimize outside the measured concentration set. The line connecting the 12 modeled conditions is only a visual guide, and a 0.2 h simulated grid does not add experimental observations between the real 6 h samples.

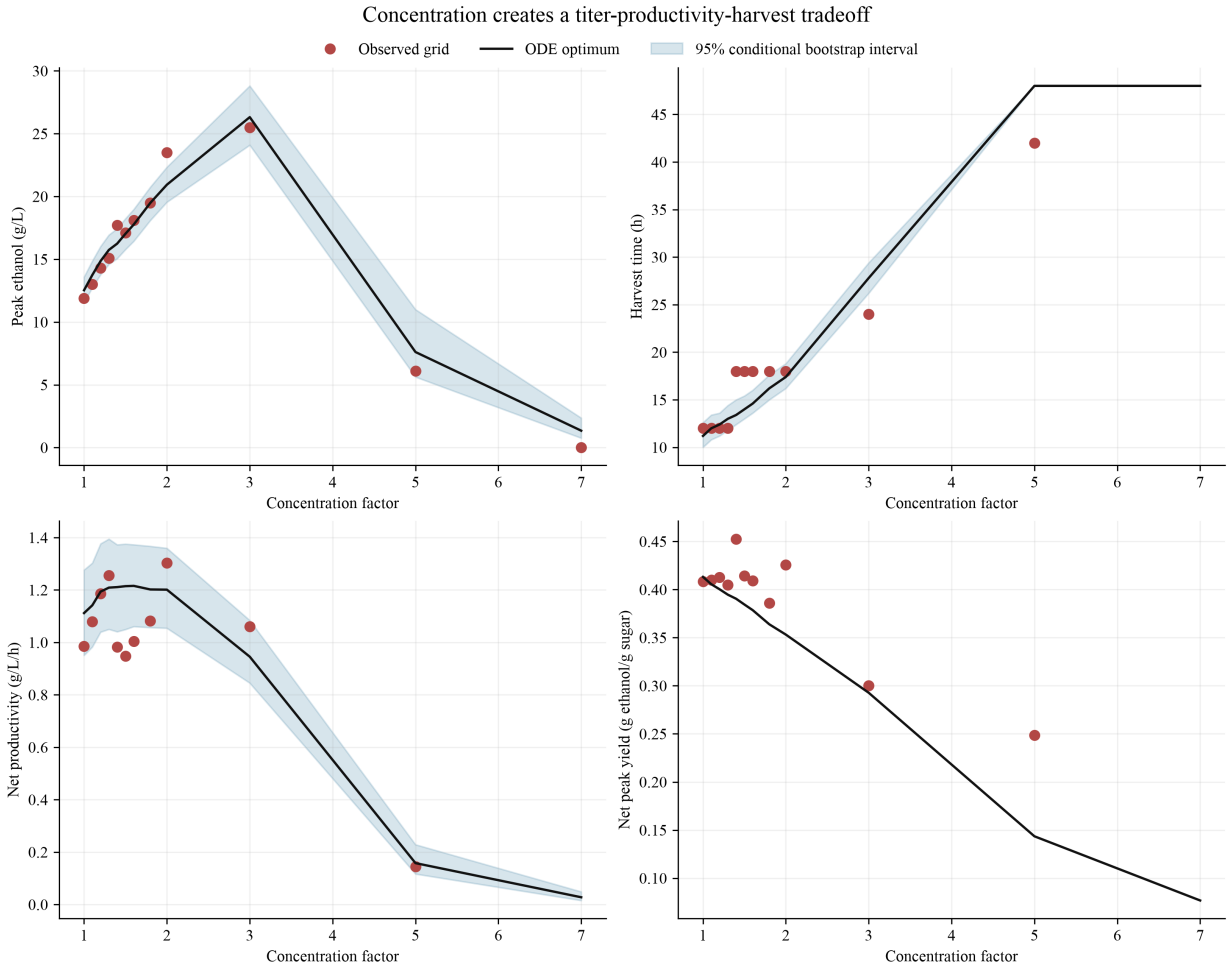

Figure 7: Observed and model-derived harvest trade-offs. Red points are values computed from the 6 h observation grid; black lines are dense-grid ODE peaks; blue bands are 95% parametric-bootstrap intervals conditional on M2, h_I=2, fixed discrepancy form, and fixed initial states. No post-inoculation ethanol peak was observed at factor 7, so its observed time, productivity, and yield are undefined. Model peaks at factors 5 and 7 occur at the 48 h boundary rather than at an interior optimum.

4.6 Rate Decomposition and Conditional Uncertainty

Figure 8 separates the model's net ethanol change into gross production and reassimilation. The modeled gross-production rates peak at 2.22 g/L/h for factor 1 and 4.38 g/L/h for factor 3. The modeled net rate crosses zero at about 11.2 h and 28.0 h as the glucose gate activates ethanol loss. At factors 5 and 7, glucose stays above the gate, so modeled reassimilation is almost zero and slow positive production continues to 48 h. This helps explain the factor-5 overprediction. These are pieces of the equation, not directly measured metabolic fluxes; real ethanol loss could include other processes.

All 200 bootstrap fits succeeded, and Figure 9 shows the conditional parameter distributions. The medians and percentile intervals are q_G,max: 2.238 [1.933, 2.561]; q_Z,max: 0.5575 [0.4699, 0.6557]; K_I: 62.09 [57.71, 67.19]; Y_E/S: 0.4657 [0.4195, 0.5136]; Y_X/S: 0.04941 [0.0444, 0.05561]; k_E: 0.006407 [0.005389, 0.007595]; Y_X/E: 0.1608 [0.1146, 0.2148]; and κ: 0.7298 [0.4299, 1.043]. The scaled Jacobian has condition number 8516, singular values from 3605.2 to 0.423, and a strongest local correlation of -0.804 between q_G,max and K_I. Broad or linked distributions show limited practical resolution. Even the narrow intervals do not remove uncertainty from fixed constants or missing processes.

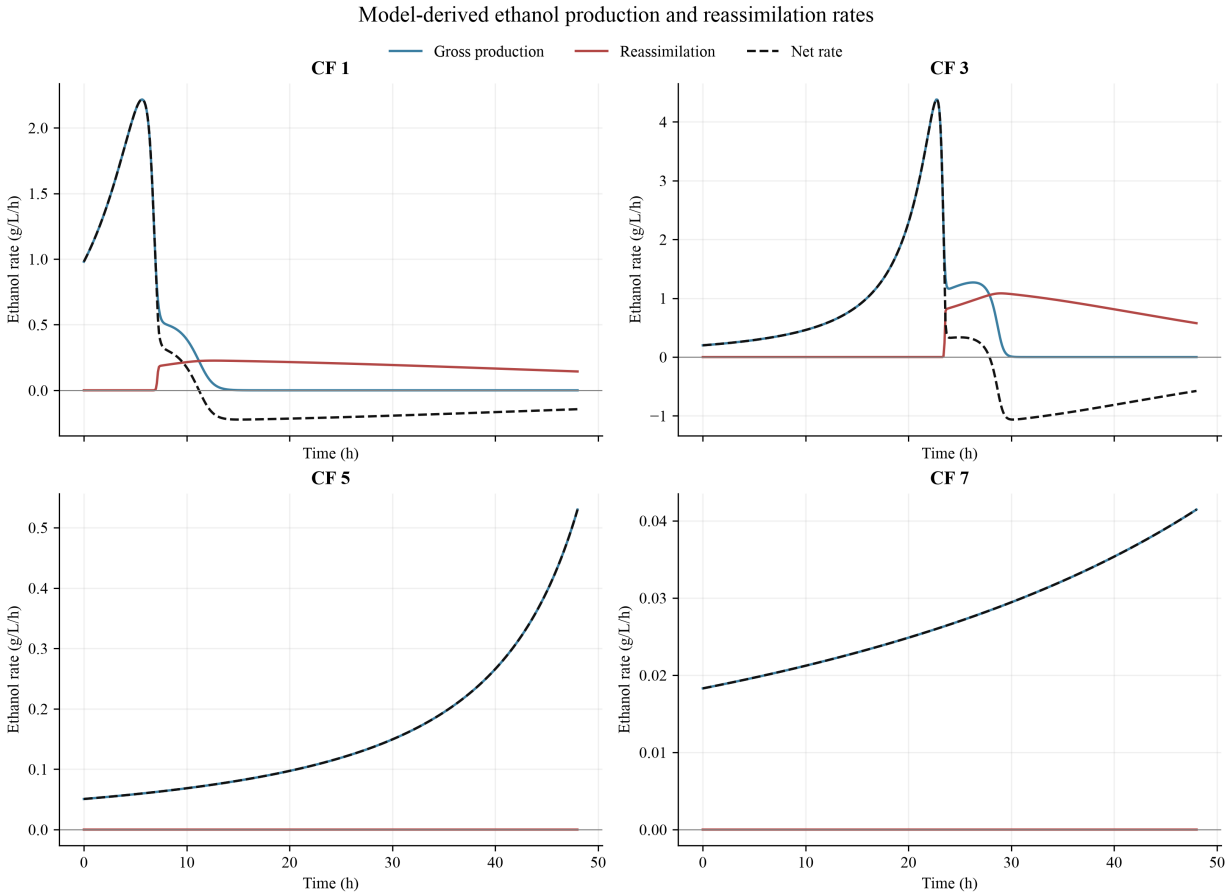

Figure 8: Model-derived ethanol-rate decomposition. Gross formation, reassimilation, and their net are latent terms of the selected equations, not independently measured metabolic fluxes. The reassimilation gate activates after modeled glucose depletion at factors 1 and 3; at factors 5 and 7 glucose remains above the fixed gate, leaving slow modeled production through the 48 h horizon.

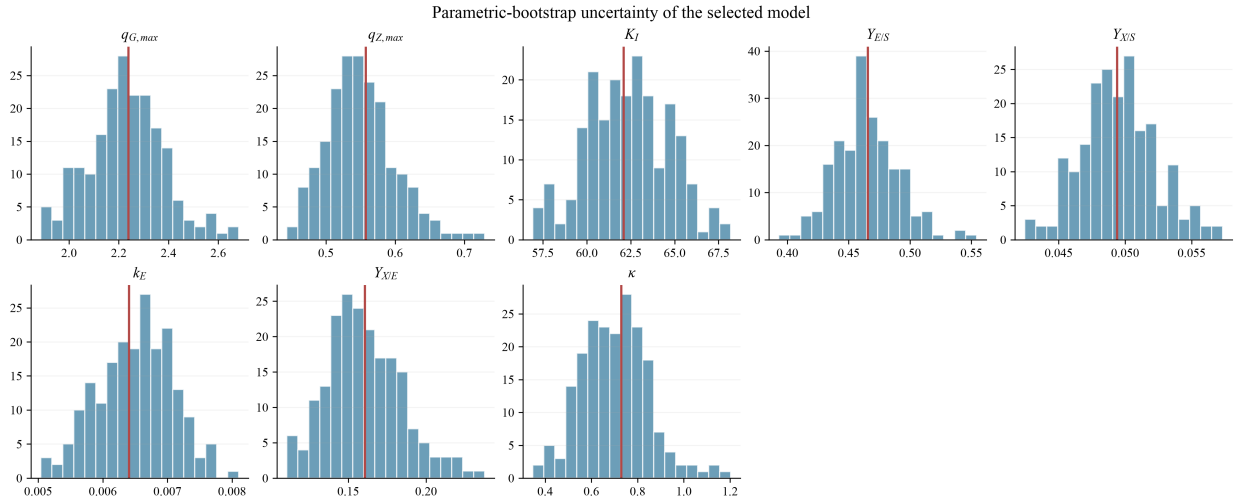

Figure 9: Conditional parameter uncertainty from 200 successful parametric-bootstrap refits. Red lines mark the selected full-data estimates. Architecture, h_I=2, fixed constants, discrepancy form, initial states, and independent Gaussian perturbations were held fixed; the histograms therefore do not include model-selection, structural, or raw-replicate uncertainty.

5 Discussion and Future Work

5.1 What Transfers

The most useful part of this project was testing whole unseen conditions instead of examining fitted curves alone. Under the equal-state score, *M_2* with *h_I=2* produced the lowest internal blocked error, 13.06% mean nRMSE. This suggests that one shared parameter set captures some common structure without using a concentration-factor correction.

The transfer is uneven, though. Low-to-moderate sugar depletion and the general rise and fall of ethanol are fairly stable. Ethanol and biomass near factors 3 and 5 are much harder. The model works across broad regimes, but it does not erase the messy boundary between them.

5.2 What the Selected Terms Mean

The nested comparison asks whether adding mathematical behavior improves prediction. Since *M_1* and *M_2* beat production-only *M_0*, the data support including a net ethanol-loss term. That is consistent with reassimilation after glucose depletion [3], but ethanol might also disappear through volatilization or unmeasured products. Likewise, *M_2* supports a concentration-dependent apparent yield inside this model, not one proven stress molecule or pathway.

The Hill exponent is best understood as a smoothness choice. Within *M_2*, *h_I=2* gave 13.06% mean blocked nRMSE, compared with 14.95% for *h_I=4* and 21.57% for *h_I=8*, even though AICc improved in the opposite direction. The fitted *K_I* also changed from 62.1 to 99.3 to 119.3 g/L. A steep curve can hug the observed transition and still predict a missing transition condition poorly. None of these exponents is a measured cooperativity coefficient.

The other parameters also depend on the chosen equations. The model has no fitted glucose-repression effect because $\rho=1$. The fixed *K_switch=0.1* g/L only opens the ethanol-use gate near glucose depletion. The apparent yields do not close a complete carbon balance. These numbers describe this model and dataset, not intrinsic constants of *C. tropicalis*.

5.3 Process Implications

More starting sugar does not guarantee more ethanol. Factor 3.0 has the largest measured titer, while factor 2.0 reaches a slightly smaller titer sooner and therefore has higher productivity. Factors 5.0 and

7.0 start with far more sugar and produce far less ethanol. An intermediate loading range looks more promising than simply filling the flask with the most concentrated hydrolysate available.

Post-peak ethanol decline also makes harvest timing important. The model can estimate when its net ethanol rate crosses zero, but the 6 h sampling grid and model-form uncertainty limit the precision of that estimate. A real harvest or feeding decision should be tested in the lab, not copied directly from a smooth simulated curve.

5.4 Limitations

The public data contain means and standard errors, not the raw replicate trajectories. I cannot reconstruct temporal correlation or batch-to-batch variation. The 6 h sampling grid also misses the exact peak time and any fast carbon-source switch. The discrepancy term τ_j helps, but the residual model is still simplified.

Concentration factor changes the whole hydrolysate. Sugar and phenolics are almost perfectly confounded, so the model cannot separate osmotic, sugar-specific, phenolic, or other matrix effects. Oxygen transfer, carbon dioxide, pH, maintenance, viability, and alternative products are missing. Exact zeros are treated as ordinary nonnegative observations rather than censored measurements.

I also used the same folds to tune exponents, compare models, and report the winning score. That makes the score selection-biased and not external validation. AICc assumes more independence than the trajectories probably contain. The bootstrap keeps the selected structure fixed, and the local Jacobian check does not prove that the parameters are unique [7]. Finally, the results apply only to this organism, feedstock, concentration range, and 48 h batch experiment.

5.5 Experiments and Computation Needed Next

A strong next experiment would vary sugar and phenolic concentrations independently. Sampling more often around the ethanol peaks-and measuring dissolved oxygen, carbon dioxide, pH, viability, volatilization, and major by-products-would help separate reassimilation from other ethanol-loss processes. Publishing raw replicate trajectories and saving an untouched concentration series or independent hydrolysate batch would finally provide a real external test.

On the computational side, profile-likelihood and structural-identifiability analyses [7], together with a hierarchical model for replicate and trajectory dependence, would improve parameter assessment. Nested blocked cross-validation could separate model selection from final performance measurement. Only after independent validation should this ODE be used for fed-batch control, harvest optimization, or prediction outside the observed grid.

6 Conclusion

6.1 Main Finding

I tested whether one four-state fermentation model could transfer across a wide concentration range. Among nine candidates, M_2 with $h\_I=2$ achieved the lowest internal model-selection score: 13.06% equal-state mean blocked nRMSE. Glucose and xylose transferred better than ethanol and biomass. The main advantage of M_2 was lower ethanol prediction error, not improvement in every state, and factors 3 and 5 exposed the biggest weakness.

6.2 Scope of the Claim

This analysis gives a reproducible, prediction-first description of the public longan-waste data. It does not prove a closed carbon balance, identify the exact cause of inhibition, measure a glucose-switch threshold, or establish external predictive accuracy. The final model is not a universal law of fermentation. It is more like a map that clearly marks both the useful roads and the places where it gets lost.

Data and Code Availability

The source measurements are available in the supplementary material from Feng et al. [1]. The reproducibility package includes the source supplement, extraction script, machine-readable data, ODE analysis, candidate scores, held-out predictions, bootstrap results, and figure code. Random seed 20260830, parameter bounds, solver tolerances, and every fixed constant are recorded. I did not generate any new experimental observations for this reanalysis.

AI Assistance Disclosure

Artificial intelligence tools were used solely to guide understanding, improve clarity, organization, and grammatical precision of the written manuscript. All research design, mathematical modeling, data analysis, code implementation, experimental validation, and scientific conclusions were developed and executed independently by the author.

References

1. Feng, J., Mahakuntha, C., Htike, S. L., Techapun, C., Phimolsiripol, Y., Rachtanapun, P., Khemacheewakul, J., Taesuwan, S., Porninta, K., Sommanee, S., Nunta, R., & Leksawasdi, N. A substrate-product switch mathematical model for the growth kinetics of ethanol metabolism from longan solid waste using *Candida tropicalis*. *Agriculture* **15**, 1472 (2025). https://doi.org/10.3390/agriculture15141472

2. Monod, J. The growth of bacterial cultures. *Annual Review of Microbiology* **3**, 371-394 (1949). https://doi.org/10.1146/annurev.mi.03.100149.002103

3. DeRisi, J. L., Iyer, V. R., & Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686 (1997). https://doi.org/10.1126/science.278.5338.680

4. Frohman, C. A., & Mira de Orduña, R. Cellular viability and kinetics of osmotic stress associated metabolites of *Saccharomyces cerevisiae* during traditional batch and fed-batch alcoholic fermentations at constant sugar concentrations. *Food Research International* **53**, 551-555 (2013). https://doi.org/10.1016/j.foodres.2013.05.020

5. Roberts, D. R., Bahn, V., Ciuti, S., et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913-929 (2017). https://doi.org/10.1111/ecog.02881

6. Hurvich, C. M., & Tsai, C. L. Regression and time series model selection in small samples. *Biometrika* **76**, 297-307 (1989). https://doi.org/10.1093/biomet/76.2.297

7. Raue, A., Kreutz, C., Maiwald, T., et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923-1929 (2009). https://doi.org/10.1093/bioinformatics/btp358