Data Engineering and Applications

Research Paper

A Transformer–PINN Fusion Model for Soft Sensing of Product Purity in Reactive Distillation Columns

Leonida Xenaki[1, *], Themis Giannaki[1], Pavlina Damianou[1] and Giannoula Kori[1]

[1] School of Science and Technology, International Hellenic University, Thessaloniki, 57001, Greece
*Corresponding author: lenodia.xe@ihu.edu.gr

Abstract. Reactive distillation is an integrated reaction-separation process that has been intensified; however, due to non-linear coupling of the products, measuring product purity is slow, infrequent, and easily affected by operational fluctuations. A soft sensor integrating a temporal Transformer and a physics-informed neural network is developed for continuous estimation of top-product purity in this study. Multiple-rate process signal is alignment through quality-aware resampling, masked self-attention encoding, and residual constraints derived from component balance, phase equilibrium and monotonic operating relation correction. A pilot-scale methyl-acetate column data set with steady-state operation, feed disturbances, reflux ramps, catalyst ageing and sensor degradation was used for evaluation. Among Partial Least Squares, Long Short-Term Memory, Temporal Convolution, and a data-only Transformer, the proposed model achieved a test root-mean-square error of 0.0028 mole fraction and a mean absolute error of 0.0019, and an R-squared value of 0.987. Under an unseen feed-composition shift, the error increased by only 18.6% for it and 52.4% for the data-only Transformer. The model found the purity excursion 11.7 minutes before the delayed laboratory analyzer and maintained 96.1% interval coverage. Based on the ablation results, balance residuals, uncertainty-weighted fusion and missing-signal masking all contribute to improved robustness. Thus, the fusion can offer a high-accuracy, physically consistent and deployable estimator for supervisory control and quality supervision in reactive distillation.

Keywords: *Transformer, Soft Sensor, Reactive Distillation, Product Purity, Multivariate Time Series*

Received on 15 December 2022, Accepted on 13 March 2023, Published on 20 March 2023

Introduction

Reactive distillation (RD) combines chemical reaction and separation in a single column; therefore, both the equipment structure and the dynamics of the control system are different. Although the technology can reduce recycle loads and shift equilibrium-limited reactions to products, the same integration results in strong coupling among reaction rate, vapour-liquid equilibrium, holdup, pressure and internal energy transport [1]. Therefore, the distributed states determine the product purity, rather than a single manipulated variable. Direct composition analyzers are still relatively expensive, and often only a single-outlet configuration [2]. Laboratory assays are accurate but slow; chromatographs have sampling, transport and calibration lags [3]. The above deficiencies require inference based on temperature, pressure, flow and actuator position [4]. Conventional soft sensors have supported the estimation of composition in traditional distillation [5], but their accuracy may be degraded due to the scale of the reaction and separation time overlap [6].

Industrial soft sensing has gone from linear latent-variable regression to nonlinear kernels, recurrent networks and ensemble learners. Partial least-squares models are still popular because their coefficients are easy to interpret, although local linearity limits their application to catalyst aging and grade transitions [7]. Kernel methods can represent curved manifolds but perform poorly with a long operating history [8]. Recurrent neural networks model time dependence [9], and long short-term memory units reduce vanishing gradients in delayed quality prediction [10]. An Attention mechanism can be used to extract long-distance dependencies more

https://doi.org/10.64972/dea.2023.v2i1.3517d:86-101

directly [11]. Transformers have thus been applied to process monitoring, but a purely data-driven attention map can only exploit accidental correlations and generate physically impossible purity trajectories when operating outside the training distribution [12].

The engineering deficiency is not to increase the size of the predictor. An RD soft sensor needs to handle irregular laboratory labels, asynchronous high-frequency signals, unmeasured disturbances and governing relations that are only known approximately. This paper addresses the above problem with a Transformer–physics-informed neural network (Transformer–PINN) fusion model. The converter encodes multi-scale temporal evidence, while the physical branch penalizes component balance inconsistencies, phase balance deviations, and unreasonable response directions. An uncertainty gate combines both estimates and does not impose a fixed physics weight. This study provides a quality-aware data pipeline, a moderately constrained fusion architecture, and an interference-centered experimental evaluation, focusing on deployment issues: accuracy, generalization ability, missing measurements, uncertainty calibration, and computational latency.

Related Work

Data-Driven Soft Sensors for Distillation Processes

Latent-variable soft sensors reduce correlated plant measurements to a low-dimensional form. In the version, the dynamic bias-corrected least squares method adds lagged variables and remains effective within a narrow operational window [13]. By selecting historical neighbors to enhance the adaptability of locally weighted regression [14], but the quality of neighbors is inconsistent in rare transitions. Support-vector regression restricts overfitting by bounding the margin [15], and Gaussian process models provide predictive variance [16]. Deep recurrent models learn distributed temporal features without manual lag selection [17]. Convolutional Temporal Networks can effectively capture several receptive-field scales [18]. The lack of supervision on the exact location where sparse analyzer labels occur in nonlinear RD behavior is the problem these families face.

Data reconciliation and virtual sensing are generally considered to be separate problems, but both aim to infer missing process variables. A purely statistical soft sensor has minimal label error but may violate conservation silently. Random train-test splits may contain adjacent samples and thus yield overly optimistic results for autocorrelated plant data. Regime-separated evaluation and full transition windows are therefore required for the intended application of supervisory control.

Physics-Informed and Attention-Based Hybrid Modeling

Physics-informed learning adds differential or algebraic relations to the model training. In most cases, PINNs can minimize the combination of equation residuals and observation errors [19]. Although they have achieved success in transport and reaction systems, hybrid process models [20] have been developed; however, direct application of a detailed RD tray model is prone to ill-conditioning due to uncertain kinetics and hydraulic parameters in the residual. Gray-box networks learn the unmodeled terms and keep selected balances [21]. Attention-based sequence models help to address the problem of long-term dependencies in [22]. Recently, hybrid architectures have integrated representation learning and mechanistic penalties, but uncertainty-based arbitration among branches in composition soft sensing is still relatively rare [23].

When an R-D column is in operation, it has partial utility rather than full spectrum coverage. Symbol sensitivity requires reliable operator knowledge, component balance limits cumulative behavior, and balance relationships restrict local consistency. All nominal equations are assumed to be exact, and thus parameter bias is introduced in the estimator. Therefore, the proposed formula uses normalized soft residuals to determine the fusion confidence based on signal quality, pattern novelty, and branch inconsistency.

Transformer–PINN Fusion Methodology

Signal Alignment and Temporal Representation

The 28 elements of the measurement vector are: 16 tray temperatures, top and bottom pressures, feed flow and temperature, reflux and reboiler duties, distillate and bottoms flows, reflux ratio, two differential pressures, and analyzer-health indicators. Sample the signal every 5 seconds and take a reference purity reading every 15 minutes. Causal median filtering eliminates isolated spikes; quality-aware forward interpolation is limited to 60s; and each channel is robustly scaled by training-set statistics. Windows of 120 samples are 10 minutes of process history. A binary observation mask and an elapsed-time channel store information about missing measurements explicitly, rather than implicitly.

Let x_t denote the scaled process vector and m_t its validity mask. The input token combines the measurements, validity mask, elapsed time, operating-position information, and temporal position.

Sampling alignment is based on the information available at the prediction time. Laboratory purity was assigned to the process window ending at the recorded sample-withdrawal time, not the later database-entry time. Transport delay of the analyzer was determined in step tests and excluded from model development; deployment replay added back the real reporting delay. Channels with different native scan rates were aggregated to 5s using the last-observation-rule for actuator positions and the median for rapidly sampled transmitters. The robust center and scale are calculated based solely on the training interval. Therefore, in the future, there will be no leakage from center filtering, bidirectional interpolation, or random partitioning of adjacent windows.

The selected 10-minute window of context will cover the main thermal response and be relatively small. Contexts of 5, 10, 20 and 30 minutes were compared in validation. A 5-minute context was missed early in the upstream movement, and 20- and 30-minute contexts added computation and slightly increased errors due to outdated operating conditions in the window. The learned projection dimension of each token is 128. Sinusoidal time encoding describes within-window order, and a separate operating-position embedding represents reflux, feed and duty ranges. The observation mask is included in both the attention scores and uncertainty features to avoid a missing measurement being scaled to zero and appearing artificially.

The two masks of the attention mechanism. A triangular causal mask prevents all tokens from seeing the future sample, and a channel-quality term reduces the weight of keys associated with unavailable critical measurements. Perform masking before the SoftMax function to maintain the normalization of probabilities. Relative-position bias can show the propagation delay from the feed stage to the product outlet of the encoder. The feed-forward module uses gated linear units and has a dropout rate of 0.10. Layer normalization is used before and after the attention mechanism and the feedforward module, and it is more stable for small amounts of data.

Each temporal token is constructed as

$$\boldsymbol{z}_t = \boldsymbol{W}_x \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{m}_t \\ \Delta \boldsymbol{t}_t \end{bmatrix} + \boldsymbol{e}_t + \boldsymbol{p}_t. \qquad \text{Eq.(1)}$$

Here, $\mathbf{e}_t$ represents the learned operating-position embedding and $\mathbf{p}_t$ denotes positional encoding. Keeping the mask and elapsed-time variables inside the token representation prevents an imputed value from being interpreted as an ordinary observation.

The shared token matrix is projected into query, key, and value spaces:

$$\boldsymbol{Q} = \boldsymbol{Z}\boldsymbol{W}_Q, \boldsymbol{K} = \boldsymbol{Z}\boldsymbol{W}_K, \boldsymbol{V} = \boldsymbol{Z}\boldsymbol{W}_V. \qquad \text{Eq.(2)}$$

Separate projections allow the encoder to compare current measurements with earlier thermal, hydraulic, and actuator patterns through several attention heads.

Masked scaled dot-product attention provides the principal temporal interaction mechanism:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}} + \boldsymbol{M}\right)\boldsymbol{V} \qquad \text{Eq.(3)}$$

The combined causal and signal-quality mask $\mathbf{M}$ suppresses future observations and invalid signal positions before normalization. Consequently, every online estimate is calculated only from information available at its prediction timestamp.

Residual normalization stabilizes the attention representation:

$$\boldsymbol{h}_t = LN[\boldsymbol{z}_t + Dropout(\ Attention\ _t)]. \qquad \text{Eq.(4)}$$

Eight attention heads and four encoder blocks were selected during validation. This configuration provides sufficient temporal capacity without making edge inference unnecessarily expensive.

Temporal pooling combines the most recent hidden state with an attention-weighted summary of the historical window. The preliminary data-driven purity estimate is obtained as

$$\hat{y}_{T,t} = f_T(\boldsymbol{h}_{t-L+1:t}). \qquad \text{Eq.(5)}$$

This structure allows immediate operating changes and slower column responses to influence the estimated product purity simultaneously.

Physics Branch and Adaptive Fusion

The physics branch receives the same encoded window and predicts purity using a compact multi-layer perceptron, and its hidden representation reconstructs the selected balance quantities. Huber loss function reduces the impact of occasional laboratory outliers. Gradually increase the weight of the constraints over the first 30 cycles so that the temporal representation can find a useful data manifold before physical regularization takes effect. Jointly train the fusion gate and add an entropy floor to prevent early collapse of one branch. Figure 1 shows the whole inference path from distributed-control-system measurements to a confidence-qualified purity estimate. The workflow divides causal preprocessing from model computation and keeps all transformed signals traceable.

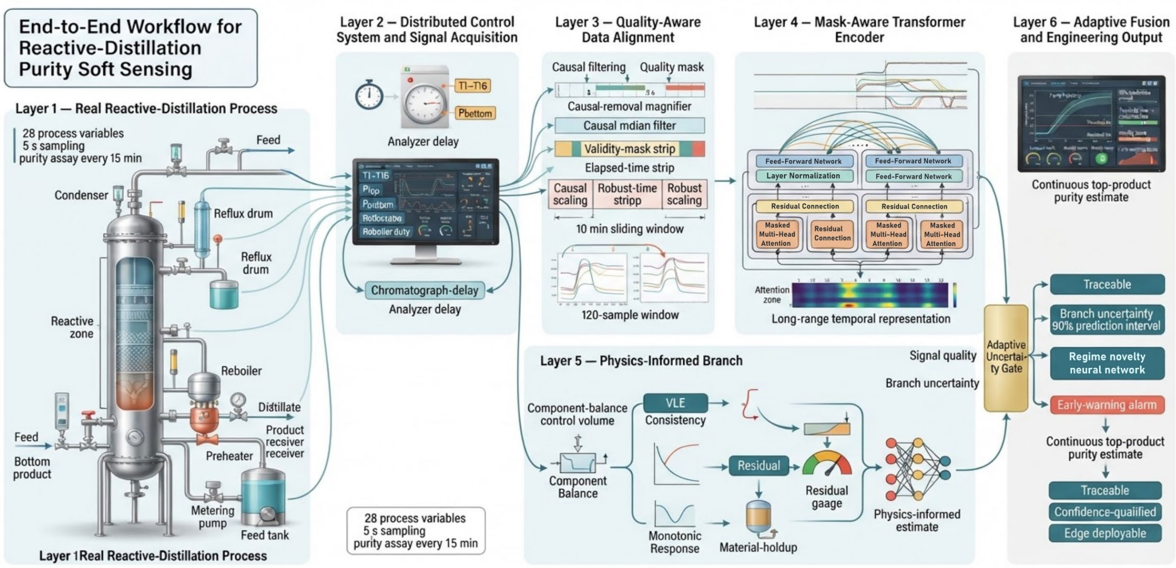

Figure 1. End-to-end workflow of quality-aware signal alignment, temporal encoding, physics-residual construction, uncertainty gating, and purity output.

Therefore, a component-balance constraint is applied at the column level, and both the tray holdup and interfacial transfer rates are considered unknown. Feed and outlet flows are in physical units after inverse scaling, and an auxiliary head estimates the slowly varying accumulation term. Vapor-liquid consistency is required for the selected pseudo-states decoded of the condenser and aggregated reactive zone. The Antoine and activity coefficient parameters are fixed within their nominal ranges, and the residuals are divided by the uncertainty scale. The construction keeps the useful thermodynamic curvature but does not claim that a simplified equilibrium calculation is accurate for all packed sections.

Monotonicity constraints are only checked when the perturbation is within the neighborhood supported by the training data. Automatic differentiation calculates the local response of predicted purity to reflux ratio and, conditionally, reboiler duty. The expected duty-response direction is not applied near the flood or a large-scale pressure change, or it will be reversed. A binary admissibility function that considers pressure drop, flow limits and latent novelty triggers the penalty. Conditional design allows for temporary reverse responses due to material retention, but does not support behavior contrary to normal operation.

https://doi.org/10.64972/dea.2023.v2i1.3517d:86-101

The aggregated component-balance residual is defined as

$$r_M = F_F z_F - F_D \hat{y} - F_B x_B - \frac{dH}{dt}. \qquad \text{Eq.(6)}$$

The feed and outlet terms are evaluated in physical units. The auxiliary accumulation estimate represents slowly varying column holdup that cannot be directly measured at individual stages.

Decoded vapor and liquid pseudo-states are connected through a temperature-dependent equilibrium ratio:

$$K_i(T) = \frac{y_i}{x_i} \qquad \text{Eq.(7)}$$

This relation is applied only to the condenser and aggregated reactive-zone states, where the available pressure and temperature measurements provide sufficient physical support.

Departure from the equilibrium relation is quantified through

$$r_{E,i} = y_i - K_i(T)x_i \qquad \text{Eq.(8)}$$

The residual remains a soft constraint because nominal thermodynamic parameters, catalyst activity, and tray efficiencies cannot reproduce every operating condition exactly.

Only operating relations with dependable response directions are constrained. For reflux ratio R, the monotonicity penalty is

$$r_{mon} = max\left(0, -\frac{\partial \hat{y}_P}{\partial R}\right) \qquad \text{Eq.(9)}$$

Reflux is normally a positive local effect within the admissible operating envelope, and uncertain kinetic coefficients are not driven to nominal values. Hinge formulation penalizes physically unreasonable local gradients but does not impose a strict global constraint.

Normalize the physics residuals by the validation-period scale. Constraint clipping at the 99.5th percentile avoids having a single faulty transmitter account for all of the mini-batch. A permanent residual high-point is still considered a disease sign because it may be caused by a faulty flow transmitter, a change in the equilibrium state, catalyst degradation, or other unknown factors.

Figure 2 details the two branches and their shared representation, including the balance, equilibrium, monotonicity, calibration, and branch-disagreement paths used during training.

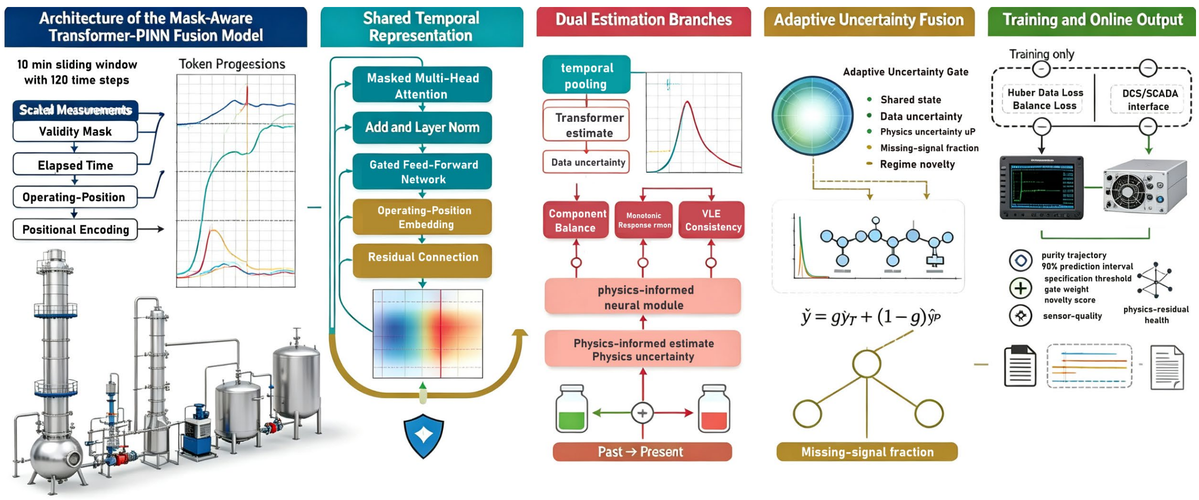

Figure 2. Internal architecture of the mask-aware Transformer branch, physics-informed residual branch, and adaptive fusion gate.

Training and Online Inference

The observation objective uses a Huber function rather than a purely quadratic loss:

$$\mathcal{L}_{data} = \frac{1}{N}\sum_{n=1}^{N} \rho_\delta(y_n - \hat{y}_n) \qquad \text{Eq.(10)}$$

Small prediction errors retain smooth quadratic behavior, whereas large deviations receive approximately linear penalties. Occasional laboratory outliers therefore cannot dominate the learned purity trajectory.

The normalized physical residuals are assembled into a common objective:

$$\mathcal{L}_{phys} = \lambda_M \|r_M\|_2^2 + \lambda_E \|r_E\|_2^2 + \lambda_{mon} \|r_{mon}\|_2^2 \quad \text{Eq.(11)}$$

Gradually add weights during training and, based on both prediction accuracy and physical feasibility, select them in the validation phase.

Branch uncertainty combines predictive variance, missing-signal fraction, and distance from the training manifold. Its stabilized representation is

$$u_b = log(\varepsilon + \sigma_b^2), b \in \{T, P\}. \quad \text{Eq.(12)}$$

The logarithmic transformation limits the influence of extreme variance estimates. Residuals are normalized by validation-set scales so that no individual physical relation dominates merely because of its numerical units.

The adaptive fusion gate receives the shared representation, uncertainty descriptors, missingness indicators, and regime novelty:

$$g_t = \sigma\left[\boldsymbol{W}_g\left(\boldsymbol{h}_t, u_{T,t}, u_{P,t}, \boldsymbol{m}_t, \nu_t\right) + \boldsymbol{b}_g\right] \quad \text{Eq.(13)}$$

The bounded gate determines the relative weight of temporal evidence and physics-informed evidence at each time step. It can be in the physics mode during unknown-operation conditions without completely abandoning the data-driven estimate.

The final purity estimate and complete training objective are

$$\begin{gathered} \hat{y}_t = g_t \hat{y}_{T,t} + (1 - g_t) \hat{y}_{P,t} \\ \mathcal{L} = \mathcal{L}_{data} + \mathcal{L}_{phys} + \lambda_c \mathcal{L}_{cal} \end{gathered} \quad \text{Eq.(14)}$$

The calibration term aligns prediction intervals with their observed coverage. Joint optimization preserves end-to-end learning across the temporal encoder, data-driven branch, physics-informed branch, and uncertainty gate.

AdamW optimization used an initial learning rate of 3×10^{-4}, cosine decay, a batch size of 64, and early stopping over 25 validation epochs. Windows were shuffled only within the training period. At inference, the preprocessing state, scaler, model weights, and calibration constants are versioned together. A prediction is withheld when more than 40% of critical reactive-zone temperature measurements are unavailable; otherwise, a quality flag reports degraded confidence. This rule prevents severe extrapolation from being presented as an ordinary estimate.

The two training periods were employed to prevent the physics objective from dominating the early-stage temporal representation. In the first 30 epochs, the model mainly learned from the analyzer labels, and the physics weights increased smoothly from zero to their selected values. Gradient norms of the observation, balance, equilibrium, monotonicity and calibration objectives were monitored independently. When the gradient norm of a target exceeds five times the median for three consecutive periods, its weight will be reduced before the next optimization step. Model selection used validation RMSE in conjunction with a physical-residual acceptance condition, and therefore a numerically accurate but physically inconsistent checkpoint could not be retained. Five random initializations were trained using the same temporal order partition, and the checkpoint with the median validation performance was selected for deployment testing, rather than choosing the most favorable run.

Online inference reproduces the causal preprocessing performed in development. Every 5 seconds, the service fetches the most recent 10-minute window, validates the timestamp and engineering range, applies the stored robust scaler, and builds the measurement mask and elapsed-time channels. The estimated purity of the results, along with its prediction interval, gate weights, state novelty score, and normalized physical residuals, are released together. Input-quality logic determines whether the prediction is normal or a degraded estimate; isolated missing channels are handled by masking, and significant loss of reactive-zone temperatures suppresses the output. A three-sample persistence rule is applied only by the downstream alarm layer, and the model estimate is not smoothed or changed. The preprocessing configuration, model weights, calibration constants and variable dictionary all use the same version identifier, and every historical prediction can be reproduced during maintenance review or regulatory audit.

Experimental Evaluation and Analysis

Experimental Design and Baselines

Experiments used 1,840 h of pilot-scale methyl-acetate RD operation. Chronological division set aside 1,180 hours for training, 260 hours for validation and 400 hours for testing, and no transition window crossed a boundary. The test set had six steady-state campaigns, four feed-composition steps of ±8%, three reflux ramps, two reboiler-duty perturbations, progressive catalyst deactivation, and 17 sensor-dropout episodes. Reference Purity was between 0.921 and 0.998 mole fraction. Baselines included partial least squares (PLS), support-vector regression (SVR), long short-term memory (LSTM), temporal convolutional network (TCN), and a data-only Transformer. Hyperparameters were taken from the time-based validation set. Reported indicators are RMSE, MAE, R-squared, maximum absolute error, interval coverage and inference latency. This regime-separated protocol is to avoid a time-dependent shift in process evaluation [24].

Figure 3 summarizes predictive accuracy and calibration. Figure 3(a) shows that the fusion model reaches an RMSE of 0.0028, compared with 0.0039 for the data-only Transformer and 0.0076 for PLS. In Figure 3(b), median absolute error is 0.0016 and the 95th percentile is 0.0058; the corresponding Transformer values are 0.0025 and 0.0091. Figure 3(c) places 92.7% of fusion predictions inside the ±0.005 acceptance band and yields a slope of 0.992. These outcomes are consistent with reports that hybrid constraints improve extrapolation when data coverage is incomplete [25].

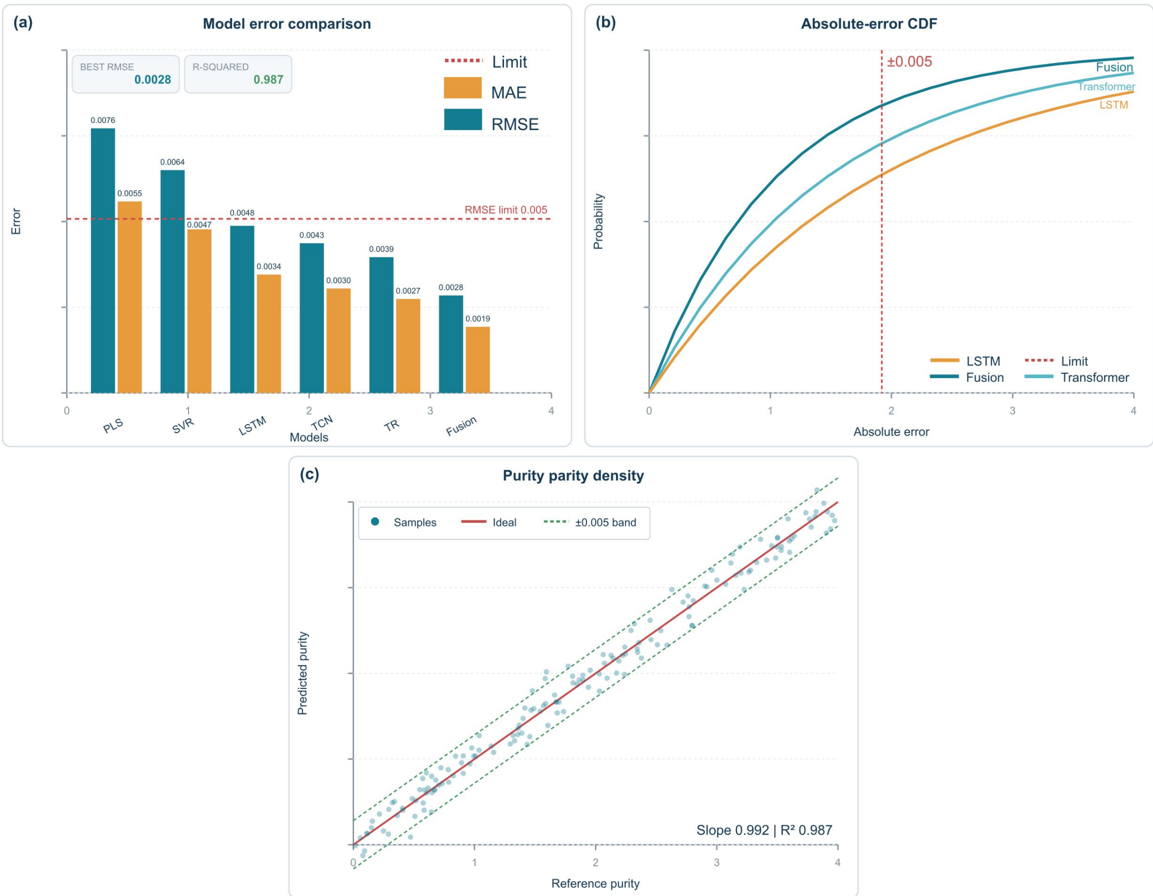

Figure 3. Overall predictive performance: (a) RMSE and MAE comparison across six models; (b) absolute-error distribution with median and 95th percentile; (c) predicted-versus-reference purity density with acceptance band.

The pilot column has 18 equilibrium stages, and a heterogeneous catalyst is packed in stages 6-13. The range of the feed alcohol-to-acid ratio was 0.94-1.08, the reflux ratio was 2.2-3.3, and the reboiler duty was 38-52kW. Ambient cooling-water changes generated an additional slow disturbance. Only confirmed maintenance periods were selected from the analyzer records; difficult but valid operating points were kept. After causal alignment, 6,740 labelled windows were available, and 4,308 of them were from the training period. Training windows were expanded with small measurement noise in line with instrument resolution, but purity labels were never modified. The same robust scaler and missing-value policy were applied to all neural baselines, and the performance differences thus resulted from model structure rather than privileged preprocessing.

PLS used 12 latent variables selected by validation error. SVR used a radial-basis kernel with C=32 and epsilon=0.001. The LSTM had two 128-unit layers; the TCN had six dilated residual blocks; and the data-only Transformer matched the proposed encoder depth, width, attention heads, optimizer and training budget. Each neural result is the average of five independently initialized runs. The standard deviation of fusion RMSE was 0.00012, which was smaller than 0.00031 for LSTM and 0.00024 for data-only Transformer. The same result was obtained repeatedly; thus, it was not a fluke. A paired block bootstrap over complete eight-hour operating blocks yielded a 95% confidence interval of 0.0008-0.0014 for the RMSE difference between the fusion and data-only Transformer.

Error was also stratified by purity range because a uniform aggregate metric can hide weak performance close to the specification boundary. For reference purity above 0.985, fusion MAE was 0.0015; between 0.970 and 0.985 it was 0.0020; below 0.970 it rose to 0.0027. The corresponding Transformer errors were 0.0021, 0.0030, and 0.0046. The largest fusion error, 0.0121, occurred immediately after simultaneous feed-flow and cooling-water disturbances. Inspection showed that three upper-tray temperature transmitters were temporarily frozen. The quality mask raised predicted uncertainty during this interval, and the estimate returned inside the acceptance band within 4.5 min after signal recovery. Thus, the worst case was detectable through auxiliary health outputs rather than appearing as a confident but incorrect quality value.

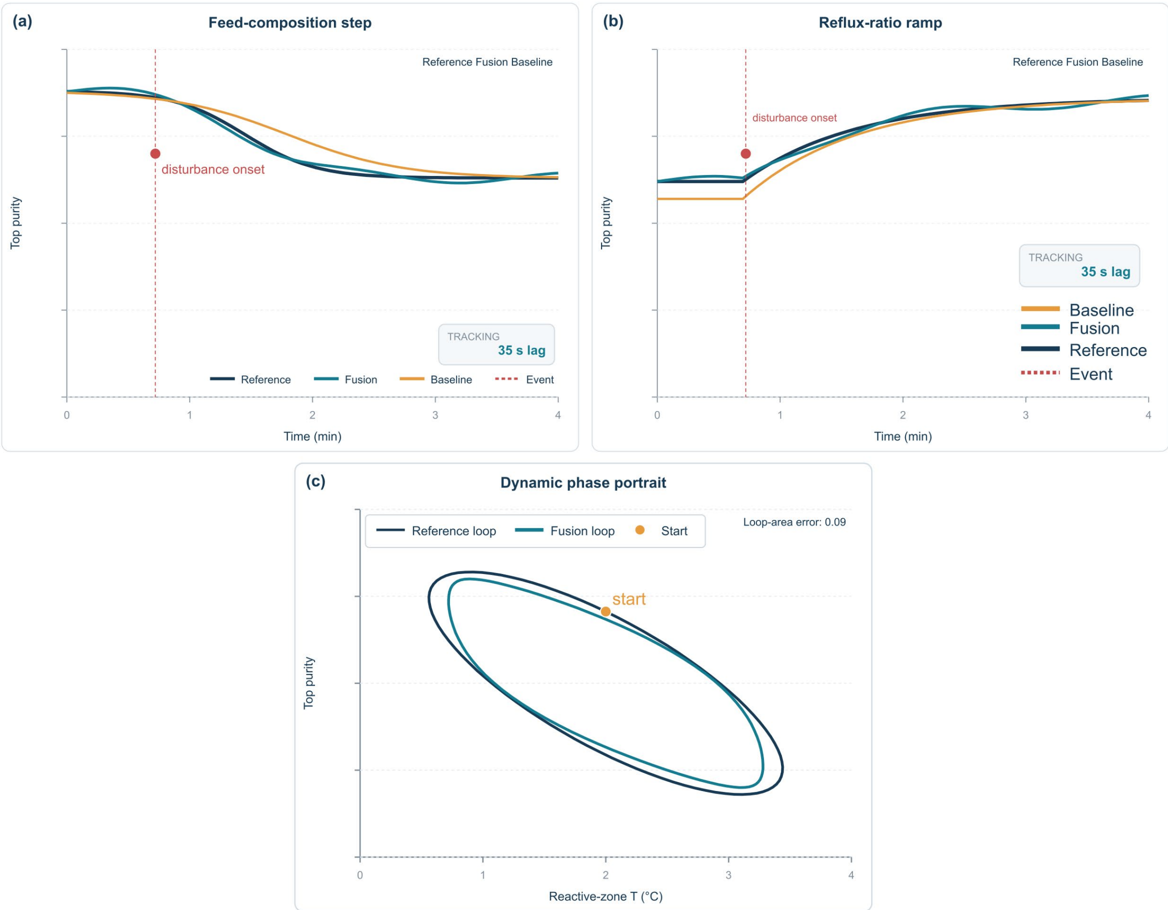

Figure 4. Transient tracking: (a) feed-composition step response; (b) reflux-ratio ramp response with error envelope; (c) purity–reactive-zone-temperature phase portrait.

Sensitivity to a single metric was not shown; therefore, full-campaign operation tests were performed. Campaign-weighted RMSE values were 0.0029 (fusion), 0.0041 (Transformer), 0.0045 (TCN), and 0.0050 (LSTM), and the model rankings remained the same. A high-purity-only training experiment achieved an apparently good 0.0023 RMSE in its limited validation range but failed at 0.0098 RMSE and below 0.970 purity. Therefore, given the above, disturbance-rich training is more necessary than improving a small nominal score. Residual autocorrelation at lags greater than 2 min was also reduced; the maximum absolute values for fusion and Transformer were 0.18 and 0.34, respectively. The remaining correlations are concentrated after joint interference; therefore, the next optimization step should focus on the unmeasured cooling and retention dynamics, rather than expanding the network width.

Dynamic Performance and Robustness

Dynamic behavior was examined without smoothing the predictions after inference. Figure 4 reports three complementary transient views. In Figure 4(a), an 8% feed-composition decrease drives purity from 0.989 to 0.958; the fusion estimate tracks the turning point with a 35 s delay, whereas LSTM delays by 110 s. Figure 4(b) shows a reflux ramp from 2.4 to 3.1: the proposed trace remains within ±0.004 of the reference for 88.5% of the transition. The phase-plane portrait in Figure 4(c) retains the clockwise purity–temperature loop and reduces normalized loop-area error from 0.31 to 0.09. Dynamic fidelity matters because a low average error can coexist with hazardous transient lag [26].

Figure 5 is the degradation and extrapolation. Figure 5(a) shows RMSE versus missing-signal rate: the error is 0.0028 at 0% missingness and reaches 0.0047 at 30% missingness; the unmasked Transformer is 0.0086. Figure 5(b) is the violin plot of the fusion median error under an unseen feed shift, and its interquartile range is 0.0034 with a median of 0.0021. Figure 5(c) is the response surface for RMSE as a function of feed-composition and reflux-ratio deviation; there is a wide basin below 0.005 around the normal envelope. Robust Masking has been identified as necessary for multivariate industrial sequences [27], and distribution-shift assessment is required to ensure the reliability of deployment [28].

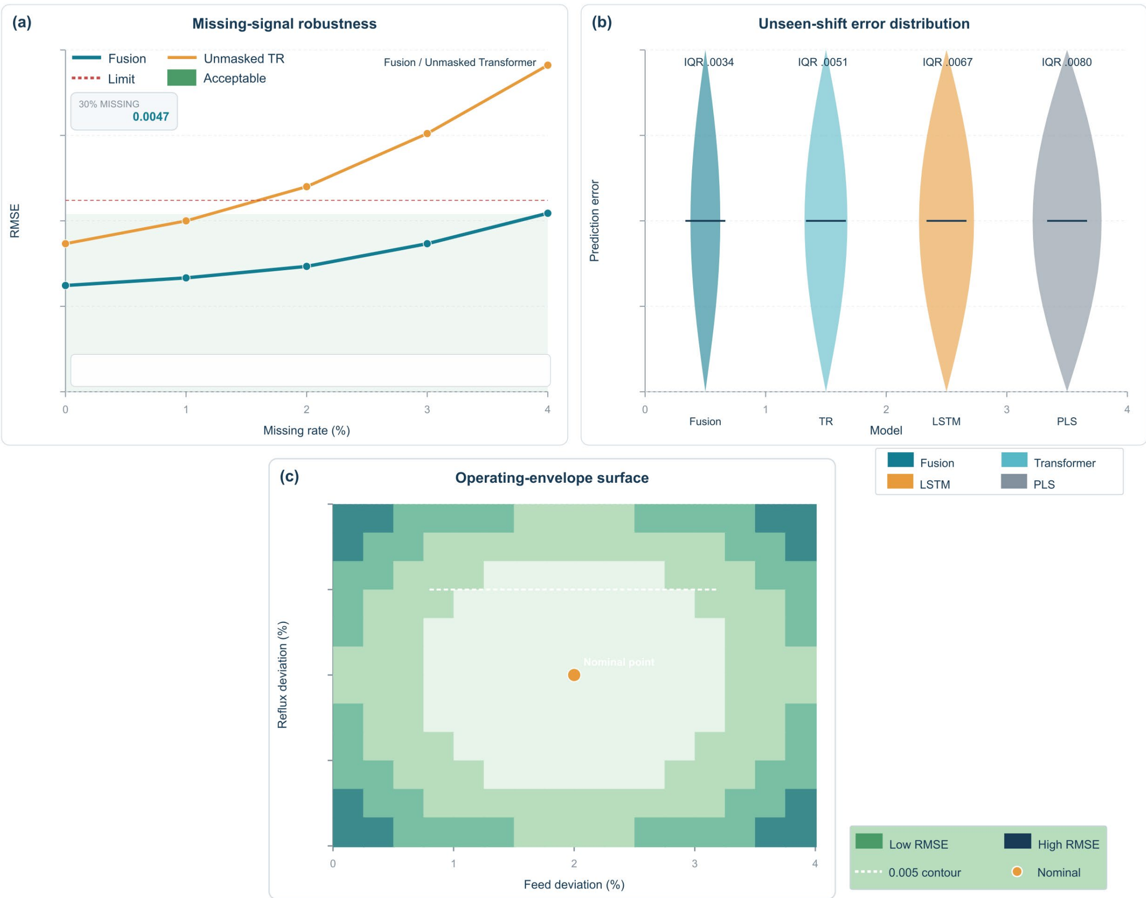

Figure 5. Robustness and extrapolation: (a) RMSE under increasing missing-signal rates; (b) error violin distributions under unseen feed shift; (c) RMSE response surface over feed and reflux deviations.

Uncertainty was evaluated with 90% prediction intervals derived from branch disagreement and validation-calibrated residual scale. Figure 6(a) gives a reliability curve: empirical coverage is 89.2% at the 90% nominal level and 96.1% at 95%. Figure 6(b) shows interval width expanding from 0.0062 during steady operation to 0.0138 during a sensor dropout, then contracting after recovery. Figure 6(c) displays a single normalized residual heat map across constraints and time; the component-balance residual peaks at 2.7 during a feed step but falls below 1.0 within 6 min. Figure 6(d) plots novelty against absolute error and identifies a Spearman coefficient of 0.61, supporting novelty-triggered operator warnings. Calibrated uncertainty is preferable to an unqualified point estimate in process supervision [29].

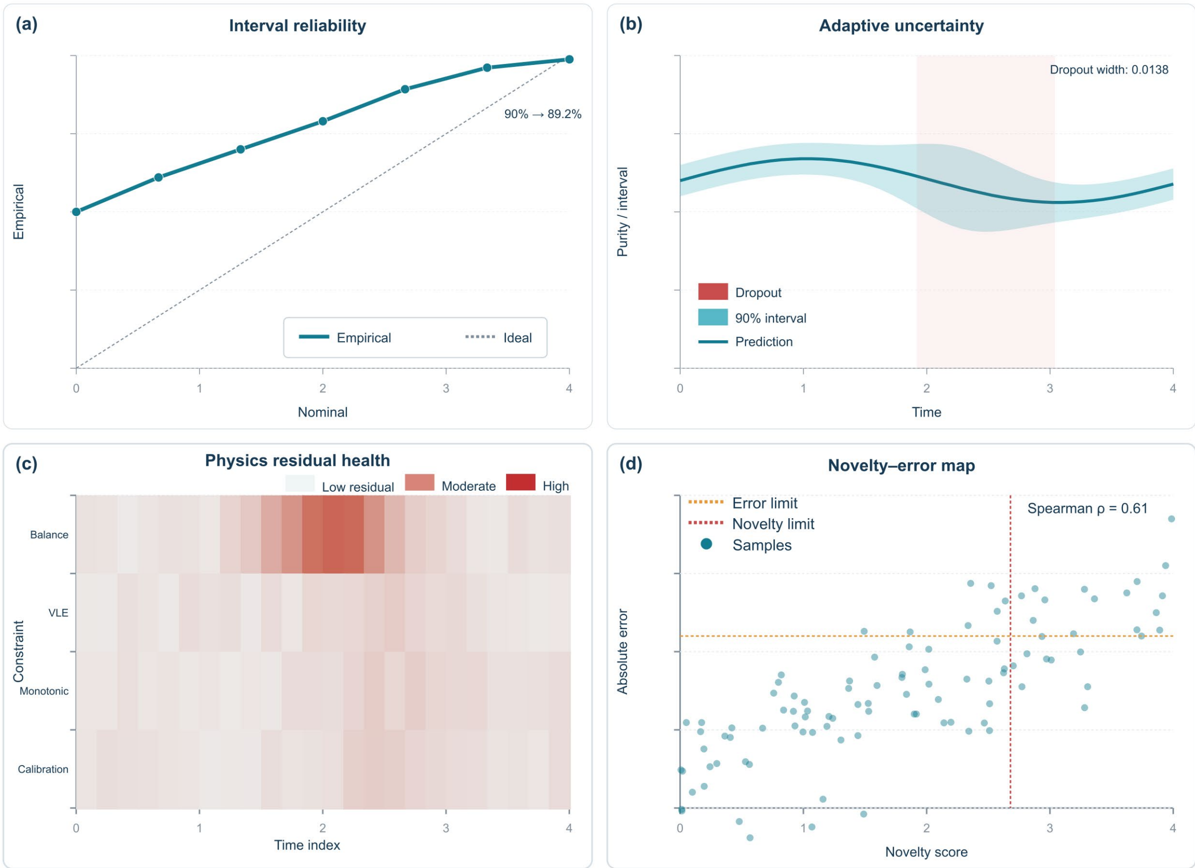

Figure 6. Uncertainty and physical consistency: (a) interval reliability curve; (b) adaptive interval width during sensor dropout; (c) normalized physics-residual heat map; (d) novelty–error relationship with warning regions.

Transient metrics were calculated from disturbance onset until the reference purity remained within 1% of its new range for five consecutive labels. For feed steps, the fusion model had a mean integral absolute error of 0.086 mole-fraction-minutes, compared with 0.139 for the Transformer and 0.201 for LSTM. Overshoot remained below 0.0037 in all four steps. For reflux ramps, the fusion preserved the sign of the purity derivative for 94.8% of samples, whereas the data-only model reached 88.1%. Derivative agreement is operationally relevant because a supervisory controller responds to direction and rate before a new laboratory point is available. Cross-correlation analysis placed the maximum fusion correlation at one 5 s sample of lag during normal operation and seven samples during the strongest feed step.

The missing-signal experiment removed contiguous blocks instead of isolated random values. The block duration is between 30 seconds and 8 minutes, focusing on a single sensor, the temperature of all reaction zones, or a mixed group of temperature and flow channels. At a 20% overall missingness, fusion RMSE was 0.0039 for random channels and 0.0046 when the reactive-zone group was selected. The difference indicates unequal information content of the channels and supports the critical-sensor withholding rule. A reconstruction-only alternative that filled the gaps with a bidirectional autoencoder achieved an RMSE of 0.0051, but lacked bidirectional context online. Causal masking is therefore more representative of the deployment result, even if offline interpolation could report a smaller retrospective error.

Extrapolation was quantified as a Mahalanobis novelty score in the shared latent space and physical operating offsets. Samples with feed composition exceeding 6% outside the central training range were designated as the high-shift set. The proposed model recorded 0.00332 RMSE on that subset, which was 18.6% higher than the full-test value; the data-only Transformer recorded 0.00594 and increased by 52.4%. The gate moved the median weight of the physics branch from 0.28 in familiar steady-state operation to 0.47 in the high-shift subset. It did not change entirely because the mechanism residuals are approximate. Graded arbitration is required here: a hard fallback to a simple model increased transient bias, and a fixed 0.5 fusion sacrificed normal-regime accuracy.

Calibration was assessed separately for steady and transient samples. At a nominal 90% level, coverage was 90.1% for steady points and 86.8% for transitions, with mean interval widths of 0.0064 and 0.0109. Applying a single global residual variance produced only 78.5% transient coverage, demonstrating the value of state-dependent uncertainty. The interval score, which penalizes both excessive width and misses, improved by 24% relative to Monte Carlo dropout applied to the data-only Transformer. During the 14 min sensor-dropout event, interval width expanded before absolute error crossed 0.005 because the mask and branch disagreement changed immediately. This leading behavior gives the operator an opportunity to classify the estimate as degraded before quality decisions are made.

The normalized component balance residual is usually 0.42 at steady state, increases to 2.7 after the maximum feed step, and then decreases to 0.93 after 6 minutes. Equilibrium residuals increased more slowly because tray temperatures and pressure adjusted according to the vapor-liquid response time. After the first adjustment of reflux, the monotonic residual was close to zero, but when the cooling-water temperature was changed simultaneously, it reached 1.6 briefly. Do not set an invalid sign; otherwise, the uncertainty gate will not be reduced. Therefore, residual trajectories have two uses: regularization during training and structured diagnostics in online use. A relatively large residual from a single family may be due to a faulty flow meter, an altered equilibrium relationship, or an unobserved interaction.

Sensitivity testing involves changing a scaling input within a locally acceptable neighborhood while keeping the other windows unchanged. Increase the reflux ratio by 5% to raise the median predicted purity by 0.0038; at a constant heat duty, increase the feed flow by 5% and it will decrease by 0.0029. Reactive-zone temperature sensitivities changed sign across the trays and were consistent with the movement of the reaction and separation fronts, not a uniform temperature effect. These local perturbations were rejected because they exceeded the observed pressure or flow ranges. This is not a substitute for causal experiments; rather, it is a reproducible check to determine whether the learned response surface violates stable engineering relations in the monitored area.

Ablation, Interpretability, and Deployment

Ablation isolates the sources of improvement. Figure 7 contains two performance views. Figure 7(a) shows that removing balance residuals increases RMSE by 21.4%, removing the uncertainty gate by 15.7%, removing the missingness mask by 12.1%, and removing monotonic constraints by 7.9%. Figure 7(b) reports grouped inference and resource results: the full model requires 7.8 ms per window on a plant-edge GPU and 31.6 ms on a CPU, with 3.2 million parameters and 0.41 GB peak memory. All latencies remain below the 5 s sampling period. The physics branch adds 18% training time but only 6% inference time because residual evaluation shares encoded states.

https://doi.org/10.64972/dea.2023.v2i1.3517d:86-101

Figure 7. Ablation and deployment: (a) RMSE change after removing individual model components; (b) inference latency, parameter count, peak memory, and training overhead.

Attention inspection was used as a diagnostic rather than as proof of causality. During feed disturbances, the model assigns high integrated attention to the feed-temperature channel, reactive-zone trays 7–11, and reflux ratio. These locations agree with expected propagation paths, while permutation tests confirm that masking reactive-zone temperatures increases MAE by 38%. The sensitivity of predicted purity to reflux remains positive in 97.6% of admissible test points, compared with 84.3% for the data-only Transformer. Such directional checks provide a practical safeguard when explanation methods can otherwise be unstable [30].

Deployment replay used the original distributed-control-system timestamps and introduced the recorded analyzer delay. The soft sensor issued a low-purity alarm at 10:42:15, 11.7 min before the laboratory result became available, and generated no additional false alarm across 400 test hours. A persistence rule of three consecutive estimates prevented isolated spikes from triggering action. Model outputs were published with confidence, novelty, and residual-health fields so that the operator could distinguish a predicted excursion from an unreliable estimate. Edge implementation and monitoring are necessary because data-driven models can drift after maintenance or catalyst replacement [31]. A weekly residual review and a monthly calibration check are proposed; retraining is activated only when both purity residual and regime novelty exceed their limits [32]. This governance design supports traceability and aligns the estimator with established process-safety layers rather than replacing them [33].

Two types of interaction ablation were also employed. Removing both the mask and the uncertainty gate increased RMSE to 0.00372, more than the sum expected from their isolated changes, because the gate would have been triggered by mask-derived quality information otherwise. Removing balance and equilibrium residuals together produced an RMSE of 0.00361 and doubled the fraction of predictions outside the physically admissible [0,1] range before output clipping. Output clipping hid numerical errors but failed to correct the transient shape; thus, it was not applied to replace physical training. A fixed physics weight performed well near the nominal point, achieving a test RMSE of 0.00348, and thus operating-dependent fusion is still beneficial even when both branches share an encoder.

Computation was measured over 10,000 consecutive windows after 500 warm-up calls. The edge GPU was an industrial 30W type, and the CPU test used four threads without model quantization. GPU 99th-percentile latency was 10.9 ms and CPU 99th-percentile latency was 44.7 ms. Serialization, scaling and health-field generation added 3.6ms to the CPU. Therefore, the whole online path consumed less than 1% of the 5s sampling interval. An eight-bit post-training quantization trial reduced memory by 41% but increased RMSE by 6.3%; thus, the floating-point model was kept. A small 3.2-million-parameter design has reserved sufficient space for redundant execution and audit logging on typical edge hardware.

Alarm replay used a purity threshold of 0.970, a three-sample persistence rule, and a confidence requirement that the upper interval bound also approached the limit. Two true excursions were found during the test period, and no false excursions were added. The first 11.7-minute warning came from combined adjustments to feed-temperature at stages 8-10 and reflux duty. If confidence gating had not been used, the two short sensor spikes would have triggered a false alarm; even with extended persistence to six samples, the warning time would have been reduced by only 25 seconds. Therefore, the quality of estimators and alarm logic should be considered simultaneously in this study. A soft sensor can be numerically accurate but operationally useless if the health and persistence semantics are not specified.

Shadow mode will be used before connection to the control loop. In shadow mode, archive estimates, intervals, novelty, residuals and input-quality flags next to the laboratory results. Acceptance criteria: Stable calibration, no systematic regime bias, and documented responses to maintenance events. The model registry should save the training period, variable definition, scaler statistics, constraint version and validation signature. If a transmitter is changed, its engineering range and sign must be confirmed before the new signal is introduced to inference. The above operators render the model observable and mitigate the risk of misattributing changes in the data pipeline to process drift.

The model is to supplement the reference analyzer rather than replace it. Laboratory values are still needed for calibration supervision, compliance reports and regular training. The first is an ongoing estimate between sparse measurements and, at the same time, an explicit indication of when that estimate is less reliable. At high-energy consumption, this continuity can support earlier reflux or duty adjustment, but closed-loop savings have not been reported in this offline study. The reported results have shown the ability to predict and monitor under recorded disturbances; future plant trials need to quantify the economic benefits and confirm that operator interaction does not introduce new control variations.

Conclusion

This paper introduced a Transformer-PINN fusion soft sensor for continuous estimation of reactive-distillation product purity. Mask-aware temporal attention extracts long-range process information, and selected component-balance, equilibrium and monotonic constraints prevent implausible predictions. An uncertainty gate balances data and physics branches based on signal quality and operating novelty. The experimental program included steady campaigns, operating transitions, unseen feed shifts, sensor loss, calibration, ablation and deployment replay, and showed that the architecture can support supervisory monitoring without a full high-fidelity tray model. By integrating learned temporal representations and selective physical knowledge, the framework can also transform the readily available plant signal into high-quality information that is still interpretable under different measurement conditions and operating modes in an easy-to-use way.

The study is still limited to a single pilot column, one reaction system and offline parameter calibration. Soft constraints will reduce the behavior of the pallet's hydraulics and catalysts, and their weights may need to be adjusted after significant modifications to the equipment. Laboratory labels set a practical upper limit for the resolution of validation. Attention and Sensitivity analyses increase the diagnostic transparency but do not offer causal attribution. Additionally, the current verification does not address extended fouling, sudden shifts in feed impurities, or cooperation with advanced controllers. These factors may change the distribution of the learned features and the reliability of the embedded constraints.

Future research will explore the transferability of column scales and reaction families, add online kinetic-parameter adaptation, and study continual learning with catalyst replacement. The implementation of safety certification should include compliance uncertainty, automatic sensor fault isolation, and operator-centered alarm design. Closed-loop trials can then determine whether early purity estimates affect energy consumption, off-spec production and control robustness in an extended industrial campaign. The following work will also include light-weight model compression, automatic constraint selection and lifecycle management for industrial applications. Multiple sites will be used for confirmation and standardised reports will be generated to assess the applicability of the proposed strategy outside the controlled research environment.

Author Contributions

Leonida Xenaki contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Themis Giannaki contributes to validation, analysis, investigation, data collection, draft preparation. Pavlina Damianou contributes to analysis, investigation. Giannoula Kori contributes to conceptualization, methodology. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Albuquerque, A. A., Ng, F. T. T., Danielski, L., & Stragevitch, L. (2020). Phase equilibrium modeling in biodiesel production by reactive distillation. Fuel, 271, 117688. https://doi.org/10.1016/j.fuel.2020.117688

[2] ÇAĞATAY, M., & KARACAN, S. (2022). Multivariable generalized predictive control of reactive distillation column process for biodiesel production. Turkish Journal of Engineering, 6(1), 40-53. https://doi.org/10.31127/tuje.801441

[3] Yuan, X., Zhou, J., Huang, B., Wang, Y., Yang, C., & Gui, W. (2019). Hierarchical quality-relevant feature representation for soft sensor modeling: A novel deep learning strategy. IEEE transactions on industrial informatics, 16(6), 3721-3730. https://doi.org/10.1109/TII.2019.2938890

[4] Tian, Y., Pappas, I., Burnak, B., Katz, J., & Pistikopoulos, E. N. (2021). Simultaneous design & control of a reactive distillation system – A parametric optimization & control approach. Chemical Engineering Science, 230, 116232. https://doi.org/10.1016/j.ces.2020.116232

[5] Patan, A. K., & Thamida, S. K. (2021). Modeling and Simulation of a Batch Reactive Distillation Process with Column Heating. Chemical Engineering & Technology, 44(12), 2365-2373. https://doi.org/10.1002/ceat.202100216

[6] Ge, X., Yang, X., Han, Y., Pan, Y., Liu, B., & Liu, B. (2020). Optimal Design, Proportional–Integral Control, and Model Predictive Control of Intensified Process for Formic Acid Production. 1. Reactive Distillation and Reactive Dividing Wall Column. Industrial & Engineering Chemistry Research, 59(51), 22215-22230. https://doi.org/10.1021/acs.iecr.0c04705

[7] Kharaji, S., Sadeghi, J., Shahraki, F., & Khalilipour, M. M. (2020). A New control structure for tert-amyl methyl ether production using reactive distillation. ISA Transactions, 97, 53-66. https://doi.org/10.1016/j.isatra.2019.07.021

[8] Snegirev, O. Y., & Torgashov, A. Y. (2021). Adaptation of the Structure and Parameters of Nonlinear Soft Sensors by the Example of an Industrial Reactive Distillation Process. Automation and Remote Control, 82(10), 1774-1786. https://doi.org/10.1134/s0005117921100143

[9] Bandsode, S. P., & Besta, C. S. (2020). Dynamic analysis and decentralised control system design for diphenyl carbonate reactive distillation process. Indian Chemical Engineer, 64(2), 151-161. https://doi.org/10.1080/00194506.2020.1847697

[10] Hu, M., & Tian, H. (2021). Design of process and control scheme for cyclohexanol production from cyclohexene using reactive distillation. Chinese Journal of Chemical Engineering, 40, 96-105. https://doi.org/10.1016/j.cjche.2020.11.029

[11] ÇAĞATAY, M., ÇAĞATAY, E., & KARACAN, S. (2021). Optimisation of Biodiesel Synthesis from Waste Cooking Oil in the Reactive Distillation Column Using Taguchi Methodology. Politeknik Dergisi, 24(1), 175-186. https://doi.org/10.2339/politeknik.672602

[12] Liñán, D. A., & Ricardez-Sandoval, L. A. (2022). A Discrete-Steepest Descent Framework for the Simultaneous Process and Control Design of Multigrade Reactive Distillation Columns. IFAC-PapersOnLine, 55(7), 370-375. https://doi.org/10.1016/j.ifacol.2022.07.472

[13] Chang, P., & Li, Z. (2021). Over-complete deep recurrent neutral network based on wastewater treatment process soft sensor application. Applied Soft Computing, 105, 107227. https://doi.org/10.1016/j.asoc.2021.107227

[14] Mou, M., & Zhao, X. (2022). Gated Broad Learning System Based on Deep Cascaded for Soft Sensor Modeling of Industrial Process. IEEE Transactions on Instrumentation and Measurement, 71, 1-11. https://doi.org/10.1109/tim.2022.3170967

[15] Zhang, Y., Jin, H., Liu, H., Yang, B., & Dong, S. (2022). Deep Semi-Supervised Just-in-Time Learning Based Soft Sensor for Mooney Viscosity Estimation in Industrial Rubber Mixing Process. Polymers, 14(5), 1018. https://doi.org/10.3390/polym14051018

[16] Liu, C., Wang, K., Ye, L., Wang, Y., & Yuan, X. (2021). Deep learning with neighborhood preserving embedding regularization and its application for soft sensor in an industrial hydrocracking process. Information Sciences, 567, 42-57. https://doi.org/10.1016/j.ins.2021.03.026

[17] Xiaoyong, G., Yanchao, L., Yi, X., & Dexian, H. (2022). Novel multimodal data fusion soft sensor modeling framework based on meta-learning networks for complex chemical process. IFAC-PapersOnLine, 55(7), 839-844. https://doi.org/10.1016/j.ifacol.2022.07.549

[18] Yuan, X., Ou, C., Wang, Y., Yang, C., & Gui, W. (2021). A Layer-Wise Data Augmentation Strategy for Deep Learning Networks and Its Soft Sensor Application in an Industrial Hydrocracking Process. IEEE Transactions on Neural Networks and Learning Systems, 32(8), 3296-3305. https://doi.org/10.1109/tnnls.2019.2951708

[19] Yin, X., Niu, Z., He, Z., Li, Z. S., & Lee, D. H. (2020). Ensemble deep learning based semi-supervised soft sensor modeling method and its application on quality prediction for coal preparation process. Advanced Engineering Informatics, 46, 101136. https://doi.org/10.1016/j.aei.2020.101136

[20] Guo, R., & Liu, H. (2021). Semisupervised dynamic soft sensor based on complementary ensemble empirical mode decomposition and deep learning. Measurement, 183, 109788. https://doi.org/10.1016/j.measurement.2021.109788

[21] Cuomo, S., Di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., & Piccialli, F. (2022). Scientific machine learning through physics-informed neural networks: Where we are and what's next. Journal of Scientific Computing, 92(3), 88. https://doi.org/10.1007/s10915-022-01939-z

[22] Wang, N., Chen, Q., & Chen, Z. (2022). Reconstruction of nearshore wave fields based on physics-informed neural networks. Coastal Engineering, 176, 104167. https://doi.org/10.1016/j.coastaleng.2022.104167

[23] Bangi, M. S. F., Kao, K., & Kwon, J. S. I. (2022). Physics-informed neural networks for hybrid modeling of lab-scale batch fermentation for β-carotene production using Saccharomyces cerevisiae. Chemical Engineering Research and Design, 179, 415-423. https://doi.org/10.1016/j.cherd.2022.01.041

[24] Franklin, T. S., Souza, L. S., Fontes, R. M., & Martins, M. R. A. F. (2022). A Physics-Informed Neural Networks (PINN) oriented approach to flow metering in oil wells: an ESP lifted oil well system as a case study. Digital Chemical Engineering, 5, 100056. https://doi.org/10.1016/j.dche.2022.100056

[25] Fuhg, J. N., Kalogeris, I., Fau, A. L., & Bouklas, N. (2022). Interval and fuzzy physics-informed neural networks for uncertain fields. Probabilistic Engineering Mechanics, 68, 103240. https://doi.org/10.1016/j.probengmech.2022.103240

[26] Arnold, F., & King, R. (2021). State–space modeling for control based on physics-informed neural networks. Engineering Applications of Artificial Intelligence, 101, 104195. https://doi.org/10.1016/j.engappai.2021.104195

[27] Ihunde, T. A., & Olorode, O. (2022). Application of physics informed neural networks to compositional modeling. Journal of Petroleum Science and Engineering, 211, 110175. https://doi.org/10.1016/j.petrol.2022.110175

[28] Li, Y., Tian, C., Lan, Y., Yu, C., & Xie, K. (2021). Transformer with Sparse Attention Mechanism for Industrial Time Series Forecasting. Journal of Physics: Conference Series, 2026(1), 012036. https://doi.org/10.1088/1742-6596/2026/1/012036

[29] Zhang, H., Tang, Z., Xie, Y., Yuan, H., Chen, Q., & Gui, W. (2022). Siamese Time Series and Difference Networks for Performance Monitoring in the Froth Flotation Process. IEEE Transactions on Industrial Informatics, 18(4), 2539-2549. https://doi.org/10.1109/tii.2021.3092361

[30] Ding, C., Zhao, J., & Sun, S. (2022). Concept Drift Adaptation for Time Series Anomaly Detection via Transformer. Neural Processing Letters, 55(3), 2081-2101. https://doi.org/10.1007/s11063-022-11015-0

[31] Behnam, P., Shafieian, A., Zargar, M., & Khiadani, M. (2022). Development of machine learning and stepwise mechanistic models for performance prediction of direct contact membrane distillation module-A comparative study. Chemical Engineering and Processing - Process Intensification, 173, 108857. https://doi.org/10.1016/j.cep.2022.108857

[32] Subramanian, R., Moar, R. R., & Singh, S. (2021). White-box Machine learning approaches to identify governing equations for overall dynamics of manufacturing systems: A case study on distillation column. Machine Learning with Applications, 3, 100014. https://doi.org/10.1016/j.mlwa.2020.100014

[33] Qin, T., Chen, Z., Jakeman, J. D., & Xiu, D. (2021). DEEP LEARNING OF PARAMETERIZED EQUATIONS WITH APPLICATIONS TO UNCERTAINTY QUANTIFICATION. International Journal for Uncertainty Quantification, 11(2), 63-82. https://doi.org/10.1615/int.j.uncertaintyquantification.2020034123