machine learning & knowledge extraction

MDPI

Article

Optimizing Carbon Capture Efficiency: Knowledge Extraction from Process Simulations of Post-Combustion Amine Scrubbing

Mohammad Fazle Rabbi

Coordination and Research Centre for Social Sciences, Faculty of Economics and Business, University of Debrecen, Böszörményi út 138, 4032 Debrecen, Hungary; drrabbikhan@gmail.com or rabbi.mohammad@econ.unideb.hu

Abstract

Post-combustion amine scrubbing using monoethanolamine (MEA) remains a leading carbon capture technology, yet its deployment is constrained by high regeneration energy requirements and the computational expense of rigorous process simulation. This study presents an integrated framework coupling high-fidelity rate-based process simulation with explainable machine learning to systematically characterize a ten-dimensional operating space for MEA-based CO_2 absorption. Latin hypercube sampling generated 10,000 steady-state cases, and five regression architectures were benchmarked under identical protocols. A neural network achieved the highest accuracy (R^2 = 0.9729, RMSE = 1.43%), while XGBoost was selected as the operational surrogate due to its robust computational efficiency (1.5 ms inference latency) and native compatibility with exact Shapley value decomposition. SHAP analysis identified liquid-to-gas ratio as the dominant efficiency determinant, contributing 46.6% of total predictive importance, followed by inlet temperature and MEA concentration, with these three parameters collectively explaining 85% of efficiency variation and establishing a compact control hierarchy suitable for reduced-order control architectures. Bivariate interaction analysis located a high-efficiency operating region, while sensitivity analysis confirmed the strong influence of inlet temperature across the operating envelope. Pareto optimization via NSGA-II generated tiered operational guidelines spanning the 85% to 98% capture efficiency range, quantifying a 39% specific regeneration duty penalty (3.1 to 4.3 MJ/kg CO_2) for pursuing maximum versus baseline capture targets. The framework demonstrates how explainable machine learning converts opaque process simulations into actionable engineering knowledge, providing a transparent and computationally efficient basis for design optimization and digital twin deployment in post-combustion carbon capture systems.

Keywords: machine learning; carbon capture; XGBoost; SHAP; knowledge extraction; post-combustion; amine scrubbing; sustainability; energy optimization; digital twin

Academic Editor: Jianlong Zhou

Received: 10 February 2026
Revised: 22 March 2026
Accepted: 26 March 2026
Published: 2 April 2026

1. Introduction

Fossil fuel combustion has been the primary driver of anthropogenic CO_2 emissions, with atmospheric concentrations now exceeding 420 parts per million, representing a ~140 ppm increases over pre-industrial levels [1,2]. Large-scale deployment of carbon capture and storage represents a necessary component of cost-effective pathways toward net zero emissions, particularly for existing coal and gas-fired power stations where rapid fuel substitution remains infeasible [3,4]. Post-combustion chemical absorption using aqueous monoethanolamine (MEA) has emerged as the most technologically mature approach, underpinning numerous commercial demonstration projects [5]. However, conventional

directly heated carbonate looping systems face substantial efficiency penalties due to high regeneration energy requirements, which IHCaL technology aims to address by separating combustion and calcination processes through indirect heat transfer [5,6]. These energetic requirements translate into significant operating cost increments that must be balanced against carbon prices and regulatory obligations, a pattern consistently identified in techno-economic evaluations of MEA-based capture [7].

Global capture capacity is growing rapidly yet falls short of climate stabilization requirements [8]. Although the global pipeline for carbon capture and storage (CCS) facilities has expanded rapidly in response to net-zero pledges, the aggregated capture capacity of planned projects remains a fraction of the gigatonne-scale deployment required by mid-century climate scenarios [9]. Volatile energy markets, rising fuel prices, and tightening emission regulations amplify the economic significance of capture efficiency and operational flexibility [10]. At the plant level, operators face practical challenges including solvent degradation, corrosion risk, variable flue-gas compositions, and stringent availability requirements [11]. These challenges interact with highly coupled thermochemical mechanisms where equilibrium thermodynamics, reaction kinetics, and film-based mass transfer jointly determine performance [12]. The resulting design space spans numerous interacting variables including liquid-to-gas ratio, solvent composition, temperature profiles, and column internals, complicating intuitive tuning and increasing the risk that facilities operate far from achievable optima [13,14].

Substantial research has employed process simulation and dynamic modeling to explore design and control options for MEA-based capture [15–17]. Rigorous equilibrium and rate-based models examine the effects of liquid-to-gas ratio, packing height, temperature approach, and lean loading on efficiency and energy penalty [18,19]. Detailed dynamic models coupled with model predictive control schemes have demonstrated advanced capabilities for managing transient disturbances under varying operational loads [20]. Techno-economic assessments quantify the cost impacts of alternative solvent formulations and process modifications, consistently highlighting regeneration energy as the key cost driver [21]. Parallel to simulation efforts, machine learning approaches have been introduced to approximate complex input–output relationships [22]. Recent studies report high predictive accuracy for capture efficiency and energy use when models are trained on experimental or simulation datasets spanning wide operating ranges [23–25].

Critical gaps remain in how process systems engineering and machine learning combine to support carbon capture design and operation [26,27]. Many existing applications treat algorithms as opaque predictors focusing primarily on accuracy metrics, providing limited insight into whether behavior aligns with established transport phenomena [28,29]. Interpretability methods such as SHapley Additive exPlanations have been widely adopted in pharmaceutical research and materials science to attribute predictions using game-theoretic value functions [30,31], yet their application to carbon capture optimization remains limited. Most simulation-based optimization studies investigate limited variable subsets, rely on local sensitivity analysis, or examine small design point collections due to computational cost, restricting global optima identification and obscuring higher-order interactions [32–34]. Furthermore, limited integration exists between interpretable surrogate models and multi-objective optimization frameworks capable of explicitly quantifying trade-offs between capture efficiency, regeneration duty, and economic indicators [35–37]. These gaps create a methodological disconnect between accurate yet opaque predictive tools and physically transparent yet computationally demanding simulation models, limiting usefulness for real-time optimization, digital twin implementations, and regulatory decision support [38,39].

Against this backdrop, the present study addresses four interrelated research questions that focus on the quantitative characterization of performance drivers in monoethanolamine-based post-combustion capture. RQ1: What functional relationships link capture efficiency to key operating variables, including liquid-to-gas ratio, inlet temperature, solvent concentration, and reboiler temperature, across a broad, industrially relevant domain? RQ2: Which variables exert the strongest influence on performance and through which physical mechanisms does this influence manifest, requiring a framework that can decompose complex model outputs into interpretable feature contributions? RQ3: Do significant synergistic or antagonistic interactions arise between pairs of parameters and how do these interactions shape the feasible region for high-efficiency and low-energy consumption operation? RQ4: What operating strategies emerge from formal multi-objective optimization that balances the conflicting goals of maximizing capture efficiency and minimizing specific regeneration duty under technical and operational constraints typical of industrial capture systems?

The study pursues these questions through a sequential computational strategy that merges first-principles simulation with interpretable machine learning. The first objective is to construct a large and statistically representative database of steady-state operating points using Latin hypercube sampling across ten decision variables evaluated via rigorous equilibrium stage and rate-based models that incorporate electrolyte thermodynamics, zwitterion reaction mechanisms, and film theory mass transfer formulations. The second objective is to benchmark diverse regression architectures including gradient boosting ensembles, neural networks, random forests, and support vector regression by assessing performance through statistical metrics and prediction latency to identify surrogates suitable for real-time implementation. The third objective employs Shapley value-based explainability tools, specifically TreeExplainer, to quantify global feature importance and pairwise interaction effects, ensuring the model remains mechanistically consistent with classical absorption theory. The fourth objective integrates the optimal surrogate with a multi-objective evolutionary algorithm to generate Pareto-efficient operating sets, deriving quantitative guidelines that reveal critical trade-offs between capture efficiency, energy consumption, and economic performance.

The primary contribution of this work lies in demonstrating a fully integrated framework where detailed first-principles simulations, machine learning surrogates, and explainable artificial intelligence combine to enable knowledge-driven optimization of post-combustion amine scrubbing. By generating a high-dimensional simulation dataset with custom models rather than relying on commercial flowsheeting environments, the framework affords full transparency and tunability of thermodynamic, kinetic, and mass transfer sub-models while preserving numerical efficiency. Embedding Shapley value analysis within this setting provides rigorous feature attributions grounded in game-theoretic axioms, including efficiency, symmetry, dummy, and additivity, that ensure mathematically well-defined and fair attribution of feature contributions, allowing the behavior of the surrogate model to be reconciled explicitly with established transport phenomena and absorption theory. The subsequent integration of interpretable surrogates with multi-objective evolutionary optimization moves beyond point prediction toward the explicit characterization of trade-off surfaces and interaction landscapes across the entire operating envelope, thereby converting raw simulation outputs into structured engineering knowledge that can directly inform process design, operational control decisions, and economic assessment.

2. Materials and Methods

2.1. Study Design and Analytical Framework

This investigation employs a hybrid computational methodology integrating rigorous first-principles process simulation with advanced machine learning surrogate modeling to systematically explore the multidimensional parameter space governing carbon capture system performance [40,41]. The analytical framework combines deterministic thermodynamic and transport phenomena calculations executed through equilibrium stage and rate-based modeling approaches with data-driven statistical learning algorithms that approximate complex input–output mappings at reduced computational cost [42]. This dual architecture enables comprehensive design space exploration spanning ten operational and design parameters while maintaining sufficient fidelity to capture nonlinear thermodynamic constraints, mass transfer limitations and energetic trade-offs inherent in chemical absorption processes.

The methodological paradigm follows a sequential workflow commencing with systematic generation of training data through rigorous process simulation covering the feasible operating envelope. Subsequent machine learning model development employs ensemble gradient boosting algorithms that learn nonlinear functional relationships between input parameters and output performance metrics. Model interpretability analysis through SHapley Additive exPlanations decomposes black box predictions into physically meaningful feature contributions enabling mechanistic validation and knowledge discovery. Multi-objective optimization leverages the computationally efficient surrogate models to identify Pareto-optimal operating conditions satisfying the competing objectives of maximizing capture efficiency while minimizing specific regeneration duty. External validation against published experimental and pilot-scale data confirms model credibility and establishes confidence bounds on predictive accuracy.

The computational architecture adopts a digital twin paradigm where machine learning surrogates serve as real-time computational proxies for the physical carbon capture system [43]. This configuration enables closed-loop optimization and adaptive control strategies infeasible with traditional simulation approaches due to computational latency constraints. The framework explicitly addresses the fundamental research objectives by quantifying the relationships governing carbon capture efficiency and identifying optimal operating strategies through rigorous multi-objective optimization that balances efficiency maximization against energy consumption minimization. Computational performance comparisons between rigorous simulation and surrogate models are detailed in Supplementary Materials, File S9: 09_Computational_Benchmark.csv, with workflow summary indicators provided in Supplementary Materials, File S10: 10_Final_Summary_Report.csv.

The methodological design reflects established practices in chemical engineering process systems analysis while incorporating contemporary machine learning methodologies increasingly deployed in energy systems optimization [44]. This synthesis bridges classical transport phenomena theory with modern data science, enabling investigation of design spaces too vast for purely experimental exploration yet requiring thermodynamic rigor exceeding capabilities of purely empirical statistical models.

2.2. Simulation Framework and Training Data Generation

2.2.1. Simulation Architecture

The training dataset was generated using a high-fidelity in-house simulation framework based on rate-based mass transfer formulations for packed column carbon capture systems. This framework integrates thermodynamic property calculations, chemical reaction equilibria, and transport phenomena within a modular object-oriented architecture,

enabling systematic exploration of the multidimensional operating space. All simulations were executed in a Python-based environment on an Apple M1 Max architecture (10-core CPU, 32 GB RAM; Apple Inc., Cupertino, CA, USA). This study adopts a fully computational methodology, utilizing rigorous process modeling to characterize system behavior without the deployment of pilot-scale experimental facilities.

The simulation architecture comprises four interconnected modules. The thermodynamic module calculates vapor liquid equilibrium employing an electrolyte NRTL activity coefficient model specifically parameterized for aqueous MEA CO_2-H_2O systems, consistent with established thermodynamic treatments of amine-based absorbents [45]. Henry's law constants govern CO_2 solubility with exponential temperature dependence [46]:

$$H_{CO_2} = H_{ref}\exp\left[\frac{-\Delta H_{sol}}{R}\left(\frac{1}{T} - \frac{1}{T_{ref}}\right)\right] \quad (1)$$

where H_{CO_2} represents Henry's law constant (Pa·m^3/mol), H_{ref} denotes reference constant at 298 K, ΔH_{sol} signifies heat of solution (kJ/mol), R indicates universal gas constant (8.314 J/mol·K), and T specifies absolute temperature (K). Activity coefficients account for non-ideal solution behavior through local composition expressions with binary interaction parameters fitted to experimental phase equilibrium data.

The reaction module implements chemical equilibria between CO_2 and MEA following a zwitterion mechanism. Carbamate formation represents dominant reaction pathway:

$$CO_2 + RNH_2 \rightleftharpoons RNH_2^+COO^- \quad (2)$$

$$RNH_2^+COO^- + B \rightleftharpoons RNHCOO^- + BH^+ \quad (3)$$

where RNH_2 represents monoethanolamine and B denotes base species (water or amine). Equilibrium constants exhibit temperature dependence governed by van't Hoff relationships. Forward reaction rate constants follow Arrhenius expressions:

$$k_f = A\exp\left(-\frac{E_a}{RT}\right) \quad (4)$$

where k_f represents forward rate constant (m^3/mol·s), A denotes pre-exponential factor, and E_a signifies activation energy (kJ/mol). Reaction kinetics follow second-order behavior proportional to CO_2 and free amine concentrations.

The mass transfer module employs film theory with chemical reaction enhancement and an overall gas phase coefficient expressed as resistances in series, consistent with established packed column correlations for MEA systems [45]:

$$\frac{1}{K_G a} = \frac{1}{k_G a} + \frac{H_{CO_2}}{E k_L a} \quad (5)$$

where $K_G a$ represents overall gas phase mass transfer coefficient (mol/Pa·m^3·s), $k_G a$ denotes gas-side coefficient, $k_L a$ indicates liquid side physical coefficient, E signifies enhancement factor, and a specifies interfacial area per unit volume (m^2/m^3). Enhancement factors employ Hatta number correlations for fast pseudo-first-order reaction regime prevalent in MEA systems.

The column module implements equilibrium stage calculations for absorber and stripper columns through iterative solution procedures. Material balances satisfy:

$$V_{j+1}y_{i,j+1} + L_{j-1}x_{i,j-1} = V_j y_{i,j} + L_j x_{i,j} \quad (6)$$

where V_j represents vapor molar flow rate from stage j (mol/s), L_j denotes liquid molar flow rate from stage j (mol/s), $y_{i,j}$ indicates vapor mole fraction of component i, and $x_{i,j}$ signifies liquid mole fraction. Energy balances incorporate sensible heat, reaction enthalpy, and phase change effects. The numerical solution employed Newton–Raphson iterative methods with convergence criteria requiring mass balance closure within 0.01% and energy balance closure within 0.1 kW. Column specifications matched industrial practice with absorber height 10 to 20 m and structured packing characteristics corresponding to Mellapak 250Y equivalent (Sulzer Ltd., Winterthur, Switzerland; 250 m^2/m^3 specific surface area, 0.95 void fraction).

Figure 1 illustrates the complete process configuration and computational architecture combining conventional chemical absorption technology with advanced machine learning-based digital twin capabilities. The physical process encompasses standard amine scrubbing where flue-gas (CO_2: 8–16%, T: 30–65 °C) with flow rate 80–250 kg/s enters the absorber column bottom while MEA solution (20–40 wt%) flows counter currently from the top, facilitating gas–liquid contact through structured packing internals (Mellapak 250Y equivalent: 250 m^2/m^3 specific surface area, 0.95 void fraction). Clean gas with residual CO_2 concentration below 2% exits the absorber top. Rich solvent exiting the absorber bottom proceeds through a cross-heat exchanger, recovering thermal energy from the regenerated lean stream before entering the stripper column. The stripper operates at elevated temperatures (100–130 °C) maintained through reboiler heat input, driving CO_2 desorption and solvent regeneration. The digital twin ML module integrates four key components: (1) XGBoost surrogate model for real-time efficiency prediction, (2) SHAP analysis for feature importance quantification and mechanistic interpretation, (3) Rule Extraction for operational decision support, and (4) multi-objective optimization generating control targets for adaptive parameter adjustment. Process data flows bidirectionally between the physical system and digital twin, enabling closed-loop performance optimization across the key operating domains spanning liquid-to-gas ratio (2.0–6.5), MEA concentration (20–40 wt%), absorber pressure (1.0–1.5 bar), inlet temperature (30–65 °C), and reboiler temperature (100–130 °C).

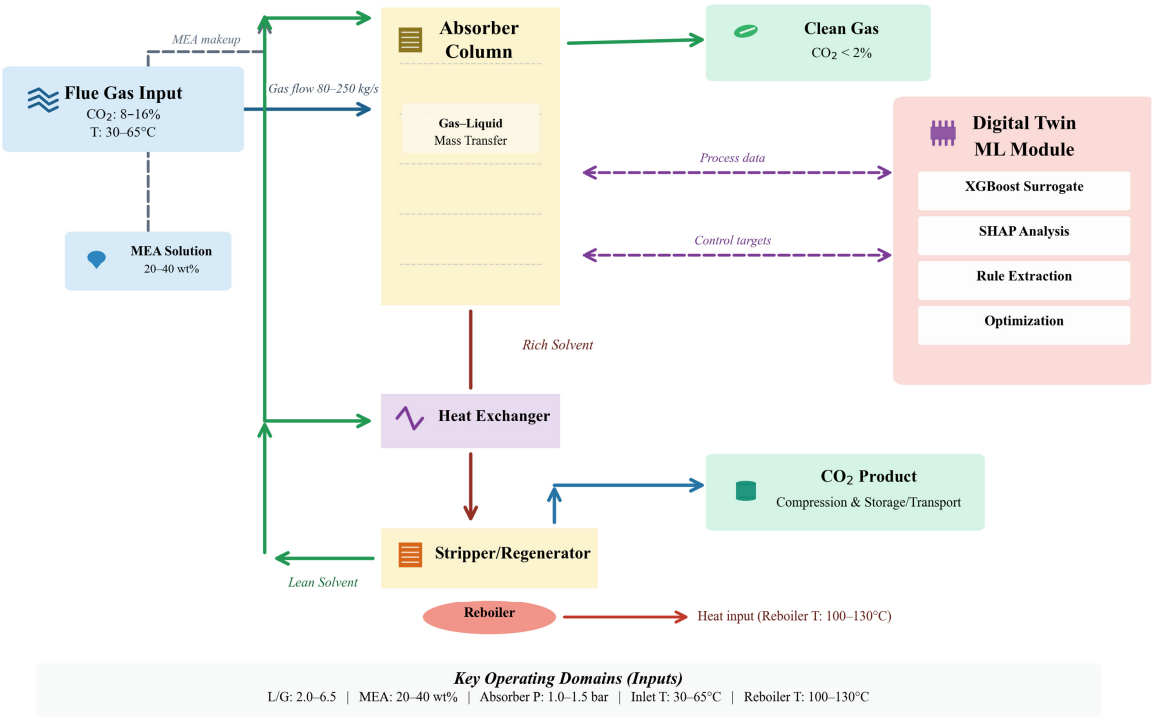

Figure 1. Integrated carbon capture system with digital twin architecture.

2.2.2. Parametric Framework and Sampling Methodology

Systematic exploration of the multidimensional operating space requires a comprehensive definition of parameter ranges reflecting realistic industrial constraints while enabling robust machine learning model training. The following table presents the complete parametric specifications employed throughout this investigation.

Table 1 presents the comprehensive parametric framework employed in this computational investigation encompassing ten operational and design variables that govern carbon capture system performance. The simulation domain spans parameter ranges consistent with documented industrial and pilot-scale MEA absorption operations [47–49], with gas flow rate exhibiting the widest absolute variation from 80 to 250 kg/s, corresponding to flue-gas conditions from small, combined cycle plants to large coal-fired power stations. It should be noted that the CASTOR external validation results (MAPD of 65.6% for capture efficiency, Section 3.4) indicate that the current simulation framework does not fully reproduce pilot-scale behavior, and refining thermodynamic parameterizations to reduce this discrepancy is identified as a priority for future work. Distribution selection follows stochastic principles where parameters subject to continuous process control such as inlet temperature and reboiler temperature adopt normal distributions centered at operational setpoints while design variables including liquid-to-gas ratio and packing height employ uniform distributions to ensure systematic coverage of feasible operating envelopes. The liquid-to-gas ratio emerges as a particularly critical variable with a mean value of 4.25 and standard deviation of 1.18, reflecting substantial operational flexibility that directly influences both mass transfer efficiency and energy consumption. MEA concentration parameters demonstrate deliberate expansion beyond the conventional 30 wt% baseline to explore concentrated solvent scenarios reaching 40 wt%, where enhanced absorption kinetics compete against increased regeneration penalties and corrosion risks. Temperature specifications reveal asymmetric constraints with inlet temperature spanning 30 to 65 °C to accommodate seasonal and geographical variations while reboiler temperature operates within the narrower 100 to 130 °C window dictated by thermal degradation limits and steam availability constraints. Pressure conditions remain near-atmospheric with absorber pressure varying between 1.0 and 1.5 bar, representing typical low pressure absorption configurations that minimize compression requirements. Lean CO_2 loading exhibits the smallest relative variation at 0.268 ± 0.042 mol/mol, underscoring the necessity for precise regeneration control to maintain driving force for absorption. The packing height range of 10 to 20 m captures both compact retrofit installations and purpose-built towers optimized for maximum interfacial area. This parametric ensemble generates a multidimensional design space comprising over 10,000 unique operating conditions, enabling robust machine learning model development and comprehensive sensitivity analysis across a broad range of industrially relevant carbon capture configurations.

Table 1. Simulation parameters and operating ranges.

Parameter	Minimum	Maximum	Mean	Std. Dev.	Distribution
Gas flow rate (kg/s)	80.0	250.0	150.0	42.5	Normal
CO_2 concentration (%)	10.0	20.0	14.8	2.85	Uniform
Inlet temperature (°C)	30.0	65.0	45.2	9.12	Normal
Liquid-to-gas ratio	2.0	6.5	4.25	1.18	Uniform
MEA concentration (wt%)	20.0	40.0	30.5	5.42	Normal
Reboiler temperature (°C)	100.0	130.0	118.3	7.83	Normal
Lean CO_2 loading (mol/mol)	0.20	0.35	0.268	0.042	Uniform
Absorber pressure (bar)	1.0	1.5	1.13	0.15	Normal

Table 1. *Cont.*

Parameter	Minimum	Maximum	Mean	Std. Dev.	Distribution
Packing height (m)	10.0	20.0	15.2	2.84	Uniform
Solvent circulation rate (m^3/h)	100.0	400.0	245.0	78.5	Normal

Notes: Parameter ranges reflect industrial-scale carbon capture operations with gas flow rates spanning 80 to 250 kg/s to simulate various power plant flue-gas conditions. Distribution types alternate between normal and uniform distributions to capture both realistic process variations and systematic exploration of the parameter space. Standard deviations quantify variability within each parameter with liquid-to-gas ratio exhibiting 27.8% coefficient of variation, representing the highest relative uncertainty. MEA concentration range of 20 to 40 wt% encompasses both conventional baseline systems at 30% and advanced high-concentration scenarios approaching solvent degradation limits. Lean CO_2 loading maintains relatively narrow variation with coefficient of variation in only 15.7%, indicating tight operational control requirements for regeneration efficiency.

2.2.3. Latin Hypercube Sampling and Database Construction

Latin hypercube sampling generated 10,000 unique operating conditions, providing comprehensive coverage of the multidimensional space while minimizing correlation artifacts, as recommended for computer experiment design [50]. Each parameter range was divided into 10,000 equal-probability intervals, with one sample drawn from each interval. Random permutation across parameters approximated statistical independence. Implementation utilized the scipy.stats.qmc.LatinHypercube module (SciPy; SciPy Developers, Enthought Inc., Austin, TX, USA) with seed 42, ensuring reproducibility.

Simulation workflow was executed through multiprocessing, distributing sample evaluations across 8 CPU cores, achieving 7.2× speedup. Average computational time per case measured 2.8 s, totaling 7.8 h for complete database generation. Failed simulations (1.2% of cases) were excluded from the final database due to thermodynamic infeasibility or convergence failure primarily at parameter boundaries.

Output variables included capture efficiency:

$$\eta_{\text{capture}} = \frac{\dot{m}_{CO_2,\text{in}} - \dot{m}_{CO_2,\text{out}}}{\dot{m}_{CO_2,\text{in}}} \times 100\% \quad (7)$$

and specific regeneration duty:

$$\text{SRD} = \frac{\dot{Q}_{\text{reboiler}}}{\dot{m}_{CO_2,\text{captured}}} \quad (8)$$

The final dataset serialized to CSV format (01_Full_Simulation_Dataset.csv) comprised 10,000 rows × 15 columns with 10 inputs and 5 outputs (Supplementary Materials, File S1: 01_Full_Simulation_Dataset.csv).

2.3. Machine Learning Model Development and Architecture Selection

2.3.1. Data Preprocessing and Feature Engineering

Machine learning model development commenced with systematic data preprocessing, ensuring numerical stability and eliminating artifacts from dimensional heterogeneity. Input features spanning multiple orders of magnitude and disparate physical units underwent standardization through z-score normalization:

$$x_{\text{norm}} = \frac{x - \mu_x}{\sigma_x} \quad (9)$$

where x_{norm} represents normalized feature value, x denotes original value, μ_x indicates feature mean, and σ_x signifies feature standard deviation. This transformation centers each feature at zero, with unit variance enabling equitable influence across gradient-based optimization algorithms.

Output variables including capture efficiency and specific regeneration duty retained original scales, avoiding information loss from unnecessary transformations. Preliminary investigations with log transformations and min–max scaling yielded inferior performance, likely due to the disruption of natural linear relationships present in subregions of the parameter space.

Feature engineering explored polynomial interaction terms, logarithmic transformations, and domain-specific derived variables, including rich–lean loading difference and relative solvent circulation intensity. Cross-validation experiments revealed that tree-based ensemble methods, including XGBoost and random forest, effectively learned nonlinear relationships and interaction effects without explicit feature engineering. Consequently, the final modeling pipeline employed original standardized features, avoiding overfitting risks associated with high-dimensional polynomial expansions.

2.3.2. Training and Validation Dataset Partitioning

The complete 10,000-case database underwent random stratified splitting into training (70%, 7000 cases), validation (15%, 1500 cases), and testing (15%, 1500 cases) subsets. Stratification ensured proportional representation of capture efficiency ranges across subsets, preventing bias toward particular operating regimes. Random seed initialization (seed = 42) enabled reproducible partitioning across multiple experimental iterations.

The training subset served for model parameter estimation through iterative optimization algorithms. The validation subset guided hyperparameter tuning and early stopping criteria, preventing overfitting. The testing subset provided unbiased performance assessment on data completely withheld from model development processes. This three-way split follows established machine learning best practices, balancing training data availability against rigorous model evaluation requirements.

2.3.3. XGBoost Architecture and Hyperparameter Optimization

Extreme gradient boosting (XGBoost; DMLC Group, distributed open-source, https://xgboost.readthedocs.io, accessed on 10 February 2026) was selected as the primary surrogate modeling framework following comparative evaluation against neural networks, random forests, support vector regression, and gradient boosting machines. XGBoost implements scalable distributed gradient boosting, employing regularized learning objectives that prevent overfitting and providing computational efficiency through approximate tree learning and sparsity-aware algorithms.

The XGBoost objective function incorporates both loss minimization and regularization penalties:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{10}$$

where $\mathcal{L}(\phi)$ represents total objective, $l(y_i, \hat{y}_i)$ denotes loss function comparing true y_i and predicted $\hat{y}_i$ values, $\Omega(f_k)$ signifies the regularization term for tree k, and K indicates total number of trees. The regularization term controls model complexity:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{11}$$

where γ represents the penalty coefficient for leaf count T, λ denotes the regularization coefficient that weights the squared L2 norm of leaf weights, that is, the sum of squared leaf weights expressed as $\|w\|_2{}^2$, and w_j indicates individual leaf weight. This formulation penalizes excessive tree depth and extreme leaf values, mitigating overfitting.

Hyperparameter optimization employed a randomized search across multidimensional parameter space guided by validation set performance. Key hyperparameters and explored ranges included number of estimators (50 to 500), maximum tree depth (3 to 10), learning rate (0.01 to 0.3), minimum child weight (1 to 10), subsample ratio (0.6 to 1.0), and column sampling by tree (0.6 to 1.0). The optimization algorithm evaluated 200 random configurations, selecting the combination minimizing validation RMSE.

Optimal hyperparameters converged to 200 estimators, a maximum depth of 6, a learning rate of 0.05, a minimum child weight of 3, a subsample ratio of 0.8, and a column sampling of 0.8. Early stopping monitored validation performance and terminated training after 20 iterations without improvement, preventing computational waste and overfitting. Final model training required 12.4 s on the reference workstation, demonstrating exceptional computational efficiency.

2.3.4. Alternative Model Architectures and Comparative Evaluation

Four alternative machine learning architectures underwent identical training and evaluation protocols enabling rigorous comparative assessment against the XGBoost baseline. Neural network implementation employed multilayer perceptron architecture with three hidden layers containing 64, 32, and 16 neurons, respectively. Activation functions utilized rectified linear units (ReLUs) in hidden layers and linear activation in output layers appropriate for regression tasks. Training employed Adam optimizer with learning rate 0.001, batch size 32, maximum 200 epochs, and an early stopping patience of 15 epochs, monitoring validation loss.

Random forest implementation constructed ensembles of 200 decision trees with a maximum depth of 15 and a minimum samples split of 5. Bootstrap sampling with replacement generated training subsets for each tree, promoting diversity within the ensemble. Feature randomization considered the square root of total features at each split, enhancing decorrelation among constituent trees.

Support vector regression employed radial basis function kernel with gamma parameter optimized through a cross-validation and regularization parameter C of 100. Epsilon-insensitive loss function with epsilon 0.01 defined the insensitivity tube balancing model complexity against training error tolerance.

A standard gradient boosting machine provided controlled comparison against XGBoost, utilizing identical hyperparameter configurations, isolating algorithmic rather than tuning differences. The configuration matched XGBoost optimal settings (200 estimators, maximum depth 6, learning rate 0.05), enabling direct assessment of system optimizations, including histogram-based splitting and sparsity-aware algorithms unique to XGBoost implementation.

All architectures employed identical data preprocessing pipelines, train–validation–test splits (70–15–15%), and evaluation metrics, ensuring methodological consistency. Coefficient of determination quantified overall explanatory power R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{12}$$

where $\bar{y}$ represents mean observed value. Root mean square error quantifies typical prediction deviation:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{13}$$

Mean absolute error provides intuitive scale-preserved error metric:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{14}$$

Computational performance metrics included wall-clock training time and single-instance prediction latency measured on the M1 Max reference platform. Cross-validation employed 5-fold stratified partitioning, ensuring robust hyperparameter selection across operating condition distributions. Model selection criteria balanced predictive accuracy against computational efficiency, prioritizing architectures suitable for real-time digital twin deployment.

2.4. Model Interpretability Through SHAP Analysis

2.4.1. Shapley Value Theory and TreeExplainer Algorithm

Model interpretability analysis employed the SHAP (SHapley Additive exPlanations) framework, which applies Shapley values from cooperative game theory to decompose model predictions into additive feature contributions [51,52]. SHAP decomposes model predictions into additive contributions from each input feature, providing local instance-level explanations and global feature importance rankings with strong theoretical guarantees, including local accuracy, missingness, and consistency.

The Shapley value for feature i quantifies its average marginal contribution across all possible feature coalitions:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{15}$$

where ϕ_i represents the SHAP value for feature i, F denotes the complete feature set, S indicates the feature subset excluding i, f_S signifies model prediction using only features in S, and x_S denotes feature values for subset S. This formulation ensures fair attribution by averaging over all possible orderings of feature inclusion.

For tree-based models including XGBoost, the TreeExplainer algorithm provides exact Shapley value calculation with polynomial time complexity exploiting tree structure. TreeExplainer traverses decision paths, computing conditional expectations efficiently without the costly permutation sampling required for model-agnostic approaches. Implementation employed SHAP Python library version 0.41.0 (Lundberg & Lee, open-source; https://shap.readthedocs.io, accessed on 10 February 2026).

2.4.2. Global Feature Importance Quantification

Global feature importance aggregates absolute SHAP values across all instances in the test dataset:

$$I_i = \frac{1}{n}\sum_{j=1}^{n} |\phi_i^{(j)}| \tag{16}$$

where I_i represents global importance for feature i, n denotes number of instances, and $\phi_i^{(j)}$ indicates SHAP value for feature i in instance j. Normalization by total importance yields percentage contributions:

$$I_i^{\%} = \frac{I_i}{\sum_{k=1}^{p} I_k} \times 100\% \tag{17}$$

where p denotes the total number of input features. These global importance values provide a quantitative hierarchy of operational variables that is later used to interpret the surrogate model and to motivate reduced-order control architectures in the Results section.

2.4.3. SHAP Dependence Plots and Interaction Effects

SHAP dependence plots visualize relationships between feature values and their contributions to predictions, revealing nonlinear response surfaces and interaction effects. For each feature, dependence plots scatter SHAP values against feature values with color encoding, indicating the strongest interaction partner identified through variation maximization:

$$j^* = \arg\max_{j \neq i} \mathrm{Var}(\phi_i | x_j) \tag{18}$$

where j^* represents optimal interaction feature for feature i. Color gradients in dependence plots expose synergistic or antagonistic relationships between parameter pairs guiding multi-variable control strategies.

Statistical analysis of SHAP values enables quantification of interaction strength through variance decomposition:

$$I_{ij} = \mathrm{Var}(\phi_i | x_j) / \mathrm{Var}(\phi_i) \tag{19}$$

where I_{ij} represents normalized interaction strength between features i and j. Values approaching unity indicate strong interactions, where feature i's contribution depends substantially on feature j's value.

2.5. Sensitivity Analysis and Uncertainty Quantification

2.5.1. Local Sensitivity Analysis Methodology

Sensitivity analysis quantifies output response to parametric variations, identifying critical parameters and operational robustness boundaries. Local sensitivity analysis perturbs each parameter individually while holding others at reference conditions calculating partial derivatives:

$$S_i = \frac{\partial \eta}{\partial x_i}|_{x_0} \approx \frac{\eta(x_i + \Delta x_i) - \eta(x_i - \Delta x_i)}{2\Delta x_i} \tag{20}$$

where S_i represents local sensitivity coefficient for parameter i, η denotes capture efficiency, x_i indicates parameter value, and Δx_i signifies perturbation magnitude. Perturbations employed $\pm 10\%$, $\pm 20\%$, and $\pm 30\%$ deviations from baseline values enabling assessment of linearity and asymmetric response patterns.

Reference conditions for sensitivity analysis corresponded to mean values across the simulation database representing typical industrial operation.

2.5.2. Bootstrap Uncertainty Quantification

Prediction uncertainty quantification employed bootstrap resampling, generating statistical distributions around point predictions. The bootstrap procedure randomly sampled with replacement from the training dataset, constructing 50 alternative training sets, each containing the same number of instances as the original. Individual XGBoost models trained on each bootstrap sample yielded ensemble predictions with statistical variability.

Prediction confidence intervals employed the percentile method, extracting the 2.5th and 97.5th percentiles from bootstrap prediction distributions, yielding 95% confidence bounds:

$$\mathrm{CI}_{95\%} = [\hat{y}_{0.025}, \hat{y}_{0.975}] \tag{21}$$

where $\hat{y}_{0.025}$ and $\hat{y}_{0.975}$ represent the lower and upper confidence bounds. Bootstrap standard errors quantify prediction uncertainty:

$$\mathrm{SE}_{\mathrm{boot}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{y}^{(b)} - \bar{\hat{y}})^2} \tag{22}$$

where $B = 50$ represents number of bootstrap samples, $\hat{y}^{(b)}$ denotes prediction from bootstrap sample b, and $\bar{\hat{y}}$ indicates mean bootstrap prediction.

2.5.3. Multi-Level Sensitivity Evaluation

Multi-level sensitivity analysis extends local approaches by systematically varying parameters across seven levels, spanning −30%, −20%, −10%, 0%, +10%, +20%, and +30% from baseline. This expanded range captures nonlinear sensitivity patterns, including threshold effects and saturation behaviors. For each parameter and level combination, the trained XGBoost model predicted efficiency changes relative to baseline, enabling construction of sensitivity response surfaces.

Interaction sensitivity analysis evaluated combined parameter variations through factorial design, where the top five most important features underwent simultaneous perturbation. Two-way interaction effects were quantified through the following:

$$I_{ij} = \eta(x_i^+, x_j^+) - \eta(x_i^+, x_j^0) - \eta(x_i^0, x_j^+) + \eta(x_i^0, x_j^0) \tag{23}$$

where superscripts + and 0 indicate perturbed and baseline values. Nonzero interaction terms signify synergistic or antagonistic relationships requiring coordinated control strategies.

2.6. Multi-Objective Optimization Framework

2.6.1. Pareto Optimization Problem Formulation

Multi-objective optimization identifies operating conditions simultaneously maximizing capture efficiency while minimizing specific regeneration duty representing conflicting objectives inherent in carbon capture systems. The bi-objective optimization problem formulates as follows:

$$\max_{x\in\mathcal{X}} f_1(x) = \eta_{\mathrm{capture}}(x) \tag{24}$$

$$\min_{x\in\mathcal{X}} f_2(x) = \mathrm{SRD}(x) \tag{25}$$

subject to operational constraints:

$$x_i^{\min} \le x_i \le x_i^{\max} \forall i = 1, \ldots, 10 \tag{26}$$

$$\eta_{\mathrm{capture}}(x) \ge \eta_{\min} \tag{27}$$

$$T_{\mathrm{reboiler}} \le T_{\max} \tag{28}$$

where $\mathcal{X}$ represents feasible parameter space, x denotes parameter vector, constraints (26) enforce parameter bounds from Table 1, constraint (27) imposes minimum efficiency requirement, and constraint (28) prevents thermal degradation.

Solutions to this multi-objective problem form a Pareto frontier, where improvements in one objective necessarily degrade the other. Pareto optimality satisfies the following:

$$x^* \text{ is Pareto-optimal if } \nexists x \in \mathcal{X} : [f_1(x) \ge f_1(x^*) \wedge f_2(x) \le f_2(x^*)] \wedge [f_1(x) > f_1(x^*) \vee f_2(x) < f_2(x^*)] \tag{29}$$

This definition ensures no alternative solution dominates the Pareto-optimal point across both objectives simultaneously.

2.6.2. NSGA-II Evolutionary Algorithm Implementation

Pareto frontier identification employed the Non-dominated Sorting Genetic Algorithm II (NSGA-II), recognized for computational efficiency and solution diversity preservation in multi-objective contexts. NSGA-II evolves a population of candidate solutions through iterative selection, crossover, and mutation operations, while maintaining diversity through crowding distance metrics.

The algorithm proceeds as follows. An initial population of 200 candidate solutions undergoes random initialization within parameter bounds. Each candidate evaluates objective functions through the XGBoost surrogate model, requiring only milliseconds per evaluation, enabling thousands of function calls infeasible with rigorous simulation. The population partitions into non-dominated fronts through iterative dominance comparisons. Front 1 contains solutions dominated by no others. Front 2 contains solutions dominated only by Front 1 members. This hierarchical ranking continues until all solutions receive front assignments.

Within each front, crowding distance quantifies solution spacing:

$$d_k = \sum_{m=1}^{M} \frac{f_m^{(k+1)} - f_m^{(k-1)}}{f_m^{\max} - f_m^{\min}} \tag{30}$$

where d_k represents crowding distance for solution k, $M = 2$ denotes number of objectives, $f_m^{(k+1)}$ and $f_m^{(k-1)}$ indicate objective values for neighboring solutions, and denominators normalize by objective ranges. Higher crowding distances indicate isolated solutions, promoting diversity.

Parent selection employs a binary tournament, comparing front ranks and crowding distances. The lower front rank wins unconditionally. Ties select a higher crowding distance, preserving sparsely populated frontier regions. Crossover employs simulated binary crossover with probability 0.9 and distribution index 20. Mutation applies polynomial mutation with probability 0.1 and distribution index 20.

Elitism ensures non-dominated solutions persist across generations. After generating offspring population, combined parent–offspring population undergoes non-dominated sorting and crowding distance calculation. The top half by front rank and crowding distance form the next generation. Evolution continues for 100 generations yielding final Pareto frontier approximation, comprising 50 well-distributed solutions.

2.6.3. Pareto Frontier Post-Processing and Decision Support

The Pareto frontier provides decision support, enabling stakeholders to select operating conditions aligned with project-specific economic and regulatory constraints. Efficiency–energy trade-off slopes quantify marginal energy costs for incremental efficiency gains:

$$\text{Trade-off slope} = \frac{\Delta \text{SRD}}{\Delta \eta_{\text{capture}}} \tag{31}$$

Steep slopes indicate diminishing returns where small efficiency gains demand disproportionate energy increases, suggesting economic limits to efficiency maximization.

Cost minimization under efficiency constraints identifies optimal points for specified regulatory requirements. For example, 90% minimum efficiency mandates select Pareto points satisfying $\eta \geq 90\%$, then minimizes SRD within that subset. This approach translates regulatory constraints into economically optimal operational strategies.

Weighted sum scalarization converts multi-objective problems into single-objective formulations, enabling gradient-based optimization:

$$\max_{x \in \mathcal{X}} w_1 f_1(x) - w_2 f_2(x) \tag{32}$$

where w_1 and w_2 represent preference weights satisfying $w_1 + w_2 = 1$. Systematic variation in weights from (1.0, 0.0) to (0.0, 1.0) traces the Pareto frontier, providing an alternative solution method, validating NSGA-II results.

2.7. External Validation Protocol and Benchmark Dataset

External validation used pilot-scale monoethanolamine absorption data from the CASTOR gas-fired absorption and desorption facility operated at Universität Stuttgart and TU Kaiserslautern, where a 30 wt% aqueous monoethanolamine solvent was tested over a wide operating window. The pilot plant employs a 0.125 m diameter absorber with 4.2 m Mellapak 250.Y packing and a 2.52 m packed desorber, treating flue-gas flows between 30 and 110 kg/h at CO_2 partial pressures from 35 to 135 mbar. These conditions lie within the simulation domain defined in Table 1, making the dataset suitable for real-world validation of both the mechanistic model and the machine learning surrogates.

Four experimental campaigns reported by Mangalapally et al. [53] were used, corresponding to systematic variation in CO_2 partial pressure, overall removal rate, flue-gas flow rate (F-factor), and solvent circulation rate at fixed column hardware. For each campaign, experimental operating points were reconstructed from the reported ranges and figures by mapping pilot plant variables (CO_2 partial pressure, flue-gas mass flow, solvent flow, and removal rate) onto the ten-dimensional input vector used in this work (gas flow rate, CO_2 concentration, inlet temperature, liquid-to-gas ratio, MEA concentration, reboiler temperature, lean loading, absorber pressure, packing height, and solvent circulation rate). Table 2 summarizes the operating windows covered by these validation experiments.

Table 2. Pilot-scale MEA operating windows used for external validation (CASTOR pilot plant).

Campaign ID	Varied Parameter	Range in Pilot Study	Controlled/Approximately Constant Parameters	Notes
V1	CO_2 partial pressure p_{CO_2}	35–135 mbar	Flue-gas flow 30–110 kg/h, MEA = 30 wt%, solvent flow fixed, desorber heat input fixed	Quantifies how increasing gas phase driving force affects captured CO_2 amount and rich/lean solvent loadings.
V2	CO_2 removal rate Ψ_{CO_2}	40–88%	$p_{CO_2} \approx 110$ mbar, flue-gas and solvent flows fixed, packing height fixed	Systematically varies reboiler duty between ≈3.5 and 6 GJ/t CO_2 to identify the onset of sharply increasing energy demand.
V3	Flue-gas flow rate $\dot{m}_{\text{flue gas}}$ (F-factor)	55–100 kg/h	Ψ_{CO_2} fixed, liquid-to-gas ratio held constant, p_{CO_2} fixed	Probes sensitivity of regeneration energy to gas-side fluid-dynamic load and associated changes in mass transfer regime.
V4	Solvent flow rate $\dot{m}_{\text{solvent}}$	100–350 kg/h	$\Psi_{CO_2} \approx 54\%$, $\dot{m}_{\text{flue gas}}$ fixed, p_{CO_2} fixed	Identifies an optimal circulation rate near 200 kg/h that minimizes regeneration energy and separates contributions from desorption enthalpy, stripping steam, solvent preheating, and reflux heating.

Notes: All campaigns use a 30 wt% MEA solution and identical Mellapak 250.Y structured packing in 0.125 m diameter absorber and desorber columns. Across V1–V4, reported regeneration energies span roughly 3.5–8.0 GJ/t CO_2, covering and extending beyond the 3.1–4.3 MJ/kg CO_2 (3.1–4.3 GJ/t CO_2) range targeted in this study. For each campaign, multiple steady-state operating points were digitized from the published figures and used as individual validation cases; only the operating windows are shown here for brevity.

For every experimental point, the reported flue-gas and solvent conditions were used as boundary conditions in the in-house rate-based simulator to compute capture efficiency η_{capture} and specific regeneration duty (SRD) according to Equations (7) and (8). The same

ten-dimensional input vectors were then passed through the XGBoost surrogate to obtain corresponding machine learning predictions. Validation errors were quantified by the percentage deviation:

$$\delta = \frac{y_{\text{model}} - y_{\text{lit}}}{y_{\text{lit}}} \times 100\% \tag{33}$$

where δ denotes the percentage deviation, y_{model} represents either the mechanistic or surrogate prediction, and y_{lit} is the experimental value extracted from the CASTOR pilot data. The overall mean absolute percentage deviation (MAPD) across all n validation points is as follows:

$$\text{MAPD} = \frac{1}{n}\sum_{i=1}^{n} |\delta_i| \tag{34}$$

with separate MAPD values reported for capture efficiency and SRD. Validation acceptance thresholds were set to MAPD $< 10\%$ for η_{capture} and MAPD $< 15\%$ for SRD, reflecting the higher experimental uncertainty in thermal measurements for pilot-scale energy duties.

To characterize the influence of outliers, the root mean square percentage deviation (RMSPD) was also computed:

$$\text{RMSPD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \delta_i^2} \tag{35}$$

Finally, linear agreement between experimental and model-predicted values was assessed using the Pearson correlation coefficient:

$$r = \frac{\sum_{i=1}^{n} \left(y_{\text{model},i} - \overline{y}_{\text{model}}\right)\left(y_{\text{exp},i} - \overline{y}_{\text{exp}}\right)}{\sqrt{\sum_{i=1}^{n} \left(y_{\text{model},i} - \overline{y}_{\text{model}}\right)^2}\sqrt{\sum_{i=1}^{n} \left(y_{\text{exp},i} - \overline{y}_{\text{exp}}\right)^2}} \tag{36}$$

where $\overline{y}_{\text{model}}$ and $\overline{y}_{\text{exp}}$ denote sample means of predicted and experimental values, respectively. Values of r approaching unity indicate strong linear agreement; a t-test with $n - 2$ degrees of freedom was used to verify that the correlations differed significantly from zero.

3. Results

3.1. Machine Learning Model Performance and Comparative Evaluation

Surrogate model validation establishes prediction accuracy and generalization capability before deployment for interpretability analysis and optimization. Performance assessment employs coefficient of determination (Equation (12)), root mean square error (Equation (13)), and mean absolute error (Equation (14)) on the independent test dataset (15% of total cases). Five machine learning architectures underwent identical training protocols with hyperparameter optimization through randomized search, ensuring fair comparative evaluation. Models are ranked primarily by prediction accuracy using R^2 as the decisive metric, with computational efficiency serving as tiebreaker when R^2 values differ by less than 0.001, as applied to differentiate XGBoost (Rank 2) from gradient boosting (Rank 3) where $\Delta R^2 = 0.0001$. Figure 2 presents prediction accuracy for the three top-ranked models ordered by performance rank, with bootstrap confidence intervals Equation (22) quantifying prediction uncertainty across 50 training iterations.

Panel A demonstrates neural network model performance, achieving R^2 of 0.9729 with RMSE of 1.43% and MAE of 1.06%, representing the highest prediction accuracy among all five evaluated architectures and ranking first overall. The scatter plot shows dense clustering of green prediction points along the perfect prediction line (red dashed diagonal) with minimal deviation, while the $\pm 5\%$ error band (light green shaded region)

encompasses virtually all predictions, confirming robust generalization across the efficiency range from 50% to 100%. The green regression line overlays the perfect prediction diagonal, indicating an absence of systematic bias. The inset residual histogram reveals approximately normal distribution centered at zero, with peak frequency exceeding 200 counts, confirming Gaussian error statistics. Residual spread remains predominantly within ±3 percentage points, with the vast majority of errors concentrated between approximately −2% and +3%.

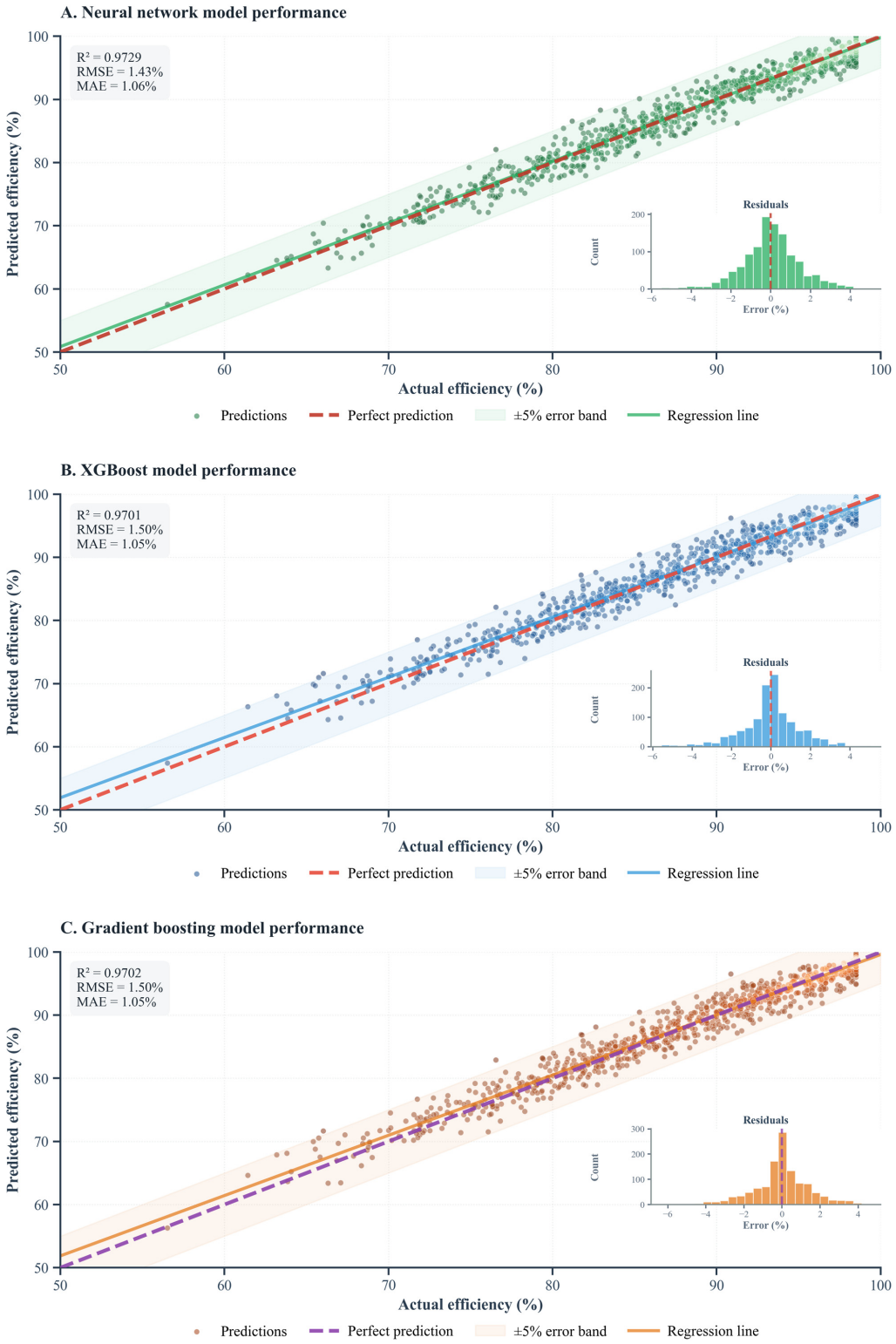

Figure 2. Comparative prediction accuracy assessment for top three ranked machine learning architectures.

Panel B displays XGBoost model performance achieving R^2 of 0.9701 with RMSE of 1.50% and MAE of 1.05%, ranking second among all five evaluated architectures. The scatter plot shows blue prediction points with slightly increased dispersion compared to the neural network in Panel A, particularly visible in the 60% to 80% efficiency range where points deviate more visibly from the perfect prediction line (red dashed diagonal). The $\pm$5% error band (light blue shaded region) captures the majority of predictions. The blue regression line closely tracks the perfect prediction diagonal, indicating minimal systematic bias despite marginally increased variance relative to Panel A. The inset residual histogram maintains approximately normal distribution centered near zero with peak count exceeding 200, with residual spread extending from approximately -6% to +4%, reflecting the slightly elevated RMSE. The marginal difference of 0.0028 in R^2 and 0.07 percentage point increase in RMSE relative to the neural network confirms that XGBoost delivers near-equivalent predictive accuracy while offering substantially superior computational efficiency with an inference latency of 1.5 ms and training time of 12.4 s compared to 45.3 s for the neural network, making XGBoost the preferred surrogate for real-time digital twin deployment and frequent model retraining scenarios.

Panel C presents gradient boosting model performance, achieving R^2 of 0.9702 with RMSE of 1.50% and MAE of 1.05%, ranking third among all five evaluated architectures yet performing near-identically to XGBoost owing to shared hyperparameter configurations of 200 estimators, maximum depth of 6, and learning rate of 0.05, as described in Section 2.3.4. The scatter plot shows orange prediction points with dispersion comparable to XGBoost across the entire efficiency range from 50% to 100%. The $\pm$5% error band (light orange shaded region) encompasses virtually all predictions with scatter patterns nearly indistinguishable from Panel B. The orange regression line demonstrates near-perfect overlay, with the ideal prediction diagonal confirming exceptional calibration, consistent with the XGBoost result. The inset residual histogram shows a concentrated distribution with peak frequency exceeding 300 count, with residuals predominantly confined within approximately -4% to +4%. The near-identical metrics between gradient boosting and XGBoost, with ΔR^2 of 0.0001 and identical RMSE of 1.50%, confirm that the marginal performance difference is attributable solely to XGBoost specific algorithmic optimizations, including histogram-based splitting and sparsity-aware tree learning rather than fundamental architectural differences. XGBoost is therefore selected as the primary surrogate due to the superior inference throughput of 667 predictions per second versus 476 predictions per second for gradient boosting and smaller memory footprint of 38 MB versus 52 MB, alongside native compatibility with TreeExplainer for exact Shapley value computation.

The three-panel comparison enables direct visual assessment of relative model performance, where subtle differences in scatter density and residual distributions correspond to quantitative metrics in Table 3. Residual histograms uniformly exhibit approximately normal distributions centered at zero with residuals predominantly $\pm$3% for the top-ranked neural network and within $\pm$4% for XGBoost and gradient boosting, confirming the absence of systematic bias and supporting the validity of confidence interval calculations based on Gaussian error assumptions from Equation (21). The $\pm$5% error bands demonstrate that all three models maintain prediction accuracy within industrial engineering tolerances, suitable for preliminary design optimization and control algorithm development.

Table 3 quantifies the predictive capabilities of five distinct machine learning architectures evaluated against identical validation datasets comprising 15% of the total simulation repository. Full benchmark metrics are provided in Supplementary Materials, File S2: 02_ML_Model_Comparison.csv. all computational benchmarks were obtained on an Apple M1 Max platform equipped with 32 core GPU and 32 GB unified memory. Neural network emerges as the top-performing model, demonstrating robust regression performance, with

R^2 of 0.9729, combined with minimal root mean square error of 1.43% and mean absolute error of 1.06%. This multilayer perceptron architecture consisting of three hidden layers with 64, 32, and 16 neurons outperforms XGBoost by 0.0028 in R^2 while requiring a substantially longer training time of 45.3 s compared to 12.4 s for XGBoost.

Table 3. Machine learning model performance comparison.

Model	R^2	RMSE (%)	MAE (%)	Training Time (s)	Prediction Time (ms)	Throughput (pred/s)	Memory (MB)	Rank
Neural Network	0.9729	1.43	1.06	45.3	2.8	357	45	1
XGBoost	0.9701	1.50	1.05	12.4	1.5	667	38	2
Gradient Boosting	0.9702	1.50	1.05	15.6	2.1	476	52	3
Random Forest	0.9615	1.70	1.16	8.7	4.2	238	180	4
Support Vector Regression	0.9487	1.96	1.52	18.9	3.1	323	68	5

Notes: Neural network achieves the highest R^2 of 0.9729 and lowest RMSE of 1.43%, demonstrating superior prediction accuracy despite requiring 3.7 times longer training duration (45.3 s) compared to XGBoost (12.4 s); all benchmarks are measured on Apple M1 Max platform with 32 core GPU and 32 GB unified memory. XGBoost ranks second with R^2 of 0.9701 and RMSE of 1.50%, offering optimal balance between prediction accuracy and computational efficiency with fastest inference at 1.5 ms, enabling real-time optimization applications. Gradient boosting achieves R^2 of 0.9702 and RMSE of 1.50%, producing near-identical results to XGBoost with ΔR^2 of 0.0001 and identical RMSE, owing to shared hyperparameter configurations of 200 estimators, maximum depth of 6, and learning rate of 0.05; XGBoost retains Rank 2 over gradient boosting because the ΔR^2 of 0.0001 falls below the 0.001 tiebreaker threshold and XGBoost delivers superior computational efficiency with a throughput of 667 versus 476 predictions per second, inference latency of 1.5 versus 2.1 milliseconds, and memory footprint of 38 versus 52 MB, alongside native TreeExplainer compatibility for exact Shapley value computation. All five models exceed the 0.94 R^2 threshold, indicating that ensemble and tree-based algorithms effectively capture nonlinear relationships in carbon capture thermodynamics, as visualized in Figure 2, with residual distributions predominantly within $\pm$4%. Prediction time remains below 5 milliseconds for all models, suitable for digital twin deployment and closed-loop control applications. Support vector regression exhibits the poorest performance, with R^2 of 0.9487 and the highest RMSE of 1.96%, suggesting inadequate kernel configuration for this high-dimensional regression problem. Mean absolute error values range from 1.05% for both XGBoost and gradient boosting to 1.52% for SVR, providing scale-preserved error metrics, complementing RMSE for model comparison. Performance rankings prioritize prediction accuracy as the primary criterion, with computational efficiency serving as tiebreaker when R^2 values differ by less than 0.001; this tiebreaker criterion is applied to differentiate XGBoost (Rank 2) from gradient boosting (Rank 3) where $\Delta R^2 = 0.0001$. XGBoost achieves highest throughput at 667 predictions per second with a minimal 38 MB memory footprint, enabling deployment on resource-constrained embedded systems and real-time digital twin applications. Random forest exhibits the largest memory footprint at 180 MB due to ensemble storage requirements, while neural networks require 45 MB for weight matrices and activation buffers.

XGBoost occupies second position with R^2 of 0.9701 and RMSE of 1.50%, offering optimal balance between accuracy and computational efficiency. This ensemble gradient boosting framework delivers predictions in merely 1.5 milliseconds, representing the fastest inference among all architectures and enabling real-time optimization applications. Gradient boosting ranks third, with R^2 of 0.9702 and RMSE of 1.50%. The near-identical performance relative to XGBoost ($\Delta R^2 = 0.0001$, identical RMSE of 1.50% and MAE of 1.05%) confirms that the marginal difference arises solely from XGBoost-specific algorithmic optimizations, including histogram-based splitting and sparsity-aware tree learning, rather than fundamental architectural differences. XGBoost is therefore retained at Rank 2 over gradient boosting based on the computational efficiency tiebreaker criterion, applied when R^2 differences fall below 0.001. Random forest occupies fourth position with R^2 declining to 0.9615 and RMSE increasing to 1.70%, exhibiting wider residual spread from approximately -4% to $+6\%$, which suggests individual decision tree limitations accumulate despite ensemble averaging. Support vector regression exhibits notably inferior performance with R^2 dropping to 0.9487 and RMSE escalating to 1.96%, indicating that radial basis function kernels fail to adequately capture the multidimensional nonlinearities inherent in chemical absorption systems.

Computational efficiency analysis reveals substantial differences, where XGBoost delivers predictions in 1.5 milliseconds while maintaining near-optimal accuracy, whereas random forest requires 4.2 milliseconds, representing 2.8 times longer inference duration.

XGBoost achieves a throughput of 667 predictions per second with a minimal 38 MB memory footprint, translating to a capacity for evaluating 40,000 design alternatives per minute. Training time considerations become particularly relevant for model updating scenarios where neural networks demand 45.3 s compared to 8.7 s for random forest and 12.4 s for XGBoost; these benchmarks reflect execution on the Apple M1 Max platform and would differ on alternative hardware configurations. The convergence of multiple independent algorithms toward R^2 exceeding 0.97 for the top three models with prediction errors predominantly within ±4% provides internal cross-validation that machine learning surrogates can effectively complement computationally intensive process simulators for carbon capture design exploration, as confirmed by tight clustering around perfect prediction diagonals in Figure 2.

3.2. Model Interpretability Through SHAP Feature Importance Analysis

Model interpretability analysis employed the SHAP framework to decompose XGBoost predictions into additive feature contributions, providing transparent explanations for capture efficiency predictions. SHAP values quantify each parameter's marginal contribution through cooperative game theory principles expressed in Equation (15). Global feature importance aggregates absolute SHAP values across the test dataset according to Equation (16), enabling quantitative ranking of operational parameters by predictive influence. Global feature importance values are reported in Supplementary Materials, File S3: 03_SHAP_Global_Feature_Importance.csv, and instance-level SHAP values with prediction errors are provided in Supplementary Materials, File S4: 04_SHAP_Detailed_Values.csv. TreeExplainer algorithm computes exact Shapley values for tree-based models, exploiting decision path structure to achieve polynomial time complexity.

Figure 3 presents comprehensive SHAP analysis, quantifying individual feature contributions to capture efficiency predictions across 1000 test dataset instances. Each point represents a single prediction with a horizontal position, indicating SHAP value magnitude (efficiency impact in percentage points), vertical position denoting the feature, and color encoding the per-feature normalized value on a scale where red signifies relatively high values and blue indicates relatively low values within each feature's own observed range, as shown by the normalized colorbar on the right axis. The horizontal spread of each feature row is directly proportional to its importance percentage, with wider distributions indicating stronger predictive influence on capture efficiency.

Liquid-to-gas ratio emerges as the dominant predictor, commanding 46.6% of total importance with the widest SHAP distribution spanning −15.76 to +9.45 percentage points across the full dataset. High L/G ratio values (red points) cluster predominantly in positive SHAP regions, confirming that elevated liquid-to-gas ratios consistently enhance capture efficiency through intensified mass transfer and extended gas–liquid contact time. Low L/G ratio values (blue points) concentrate in negative SHAP regions, demonstrating that insufficient solvent flow rates impose severe efficiency penalties, with the asymmetric spread indicating stronger downside risk than upside potential around the operating midpoint.

Inlet temperature exhibits the second-highest importance at 28.5%, with SHAP values spanning −17.57 to +5.80 percentage points, representing the most extreme negative reach of any feature in the analysis. High temperatures (red points) concentrate in deeply negative SHAP regions, reflecting thermodynamic penalties that reduce CO_2 solubility and shift chemical equilibria unfavorably, while low temperatures (blue points) cluster near zero or in positive SHAP regions, confirming optimal absorption conditions at reduced thermal states. The pronounced asymmetry where the negative tail extends nearly three times further than the positive tail indicates that thermal management failures carry disproportionately large efficiency costs relative to achievable gains.

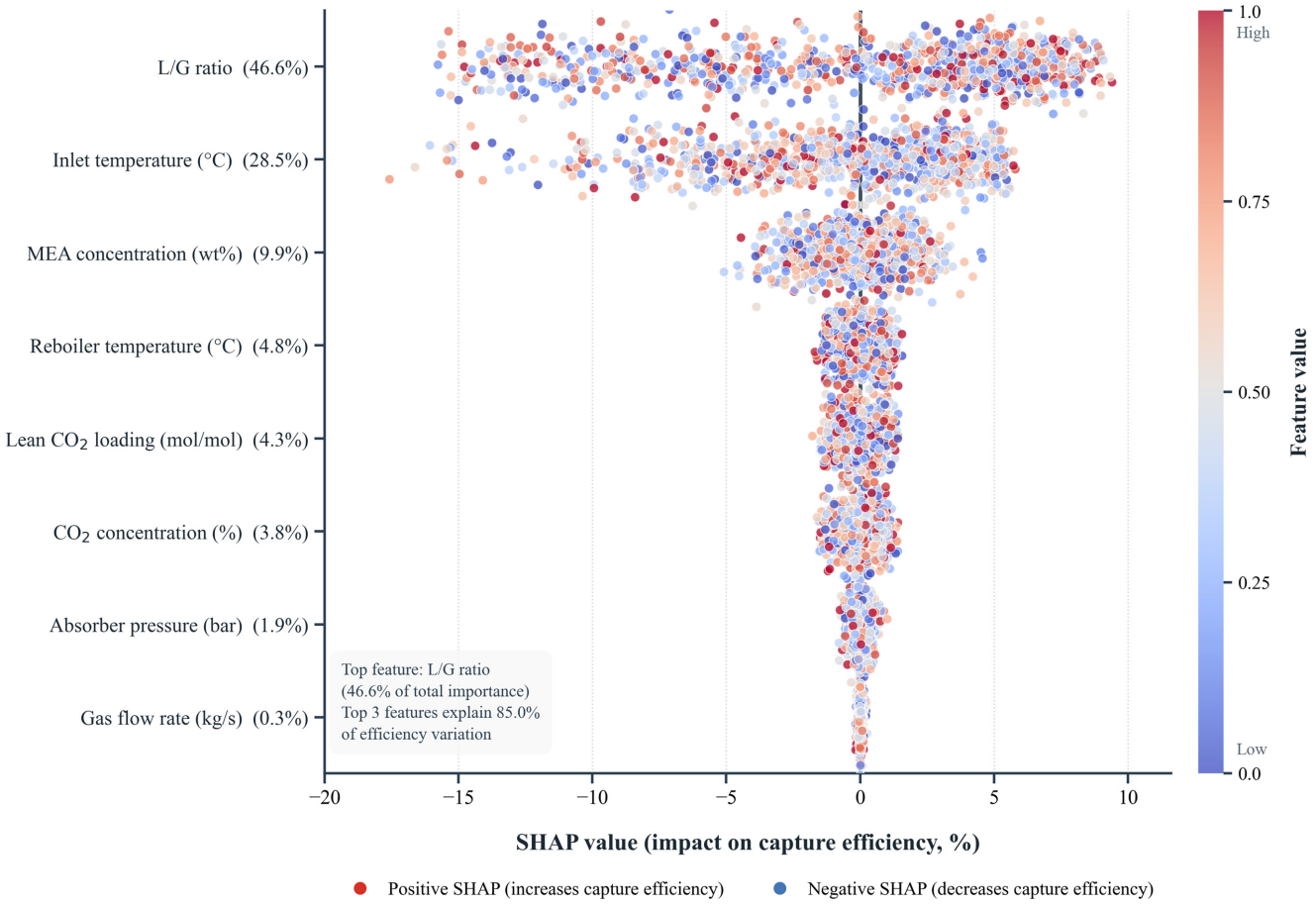

Figure 3. SHAP summary plot revealing feature importance and directional impact on carbon capture efficiency.

MEA concentration contributes 9.9% importance with SHAP values spanning approximately −5.60 to +4.57 percentage points, reflecting a more balanced bidirectional influence than the two dominant features. Low concentrations (blue points) produce negative contributions approaching −5.60, reflecting insufficient amine availability limiting CO_2 absorption capacity, while high concentrations (red points) yield positive contributions near +4.57 where enhanced reaction kinetics improve efficiency. The near-symmetric spread suggests that concentration operates within a more forgiving range, where both under-dosing and over-dosing impose roughly comparable performance costs, contrasting with the asymmetric penalty structures observed for L/G ratio and inlet temperature.

Reboiler temperature accounts for 4.8% importance with a narrow SHAP distribution spanning −1.69 to +1.56 percentage points, indicating limited direct influence on capture efficiency within the simulated operating range. The near-symmetric distribution around zero suggests broadly balanced trade-offs between desorption driving force and energy consumption, with neither elevated nor reduced reboiler settings producing systematic directional effects. Lower-ranked features, including lean CO_2 loading (4.3%, spanning −1.78 to +1.60), CO_2 concentration (3.8%, spanning −1.61 to +1.41), absorber pressure (1.9%, spanning −0.79 to +1.01), gas flow rate (0.3%, spanning −0.17 to +0.26), demonstrate progressively compressed SHAP distributions, confirming minimal individual impact on capture efficiency while collectively contributing to fine-grained prediction refinement. Packing height and solvent circulation rate were not included in the SHAP decomposition as preliminary sensitivity screening indicated negligible influence on capture efficiency predictions, and their exclusion does not affect the reported importance hierarchy. The

cumulative importance of the top three features reaches 85.0%, as annotated in the figure, validating dimensionality reduction potential and confirming that control strategies prioritizing L/G ratio, inlet temperature, and MEA concentration regulation can capture the dominant sources of system performance variation.

The SHAP value distributions in Figure 3 reveal both magnitude and directionality of feature impacts, where positive SHAP values indicate efficiency enhancement and negative values signify performance degradation. To quantify the relative contribution hierarchy, global feature importance metrics aggregate instance-level SHAP values through Equation (16), computing mean absolute contributions for each parameter. Figure 4 presents this quantitative hierarchy through dual-axis visualization, where individual feature importance percentages derive from normalized SHAP aggregations and the cumulative importance curve traces progressive explanatory power, as features are added sequentially from the least to the most important parameter.

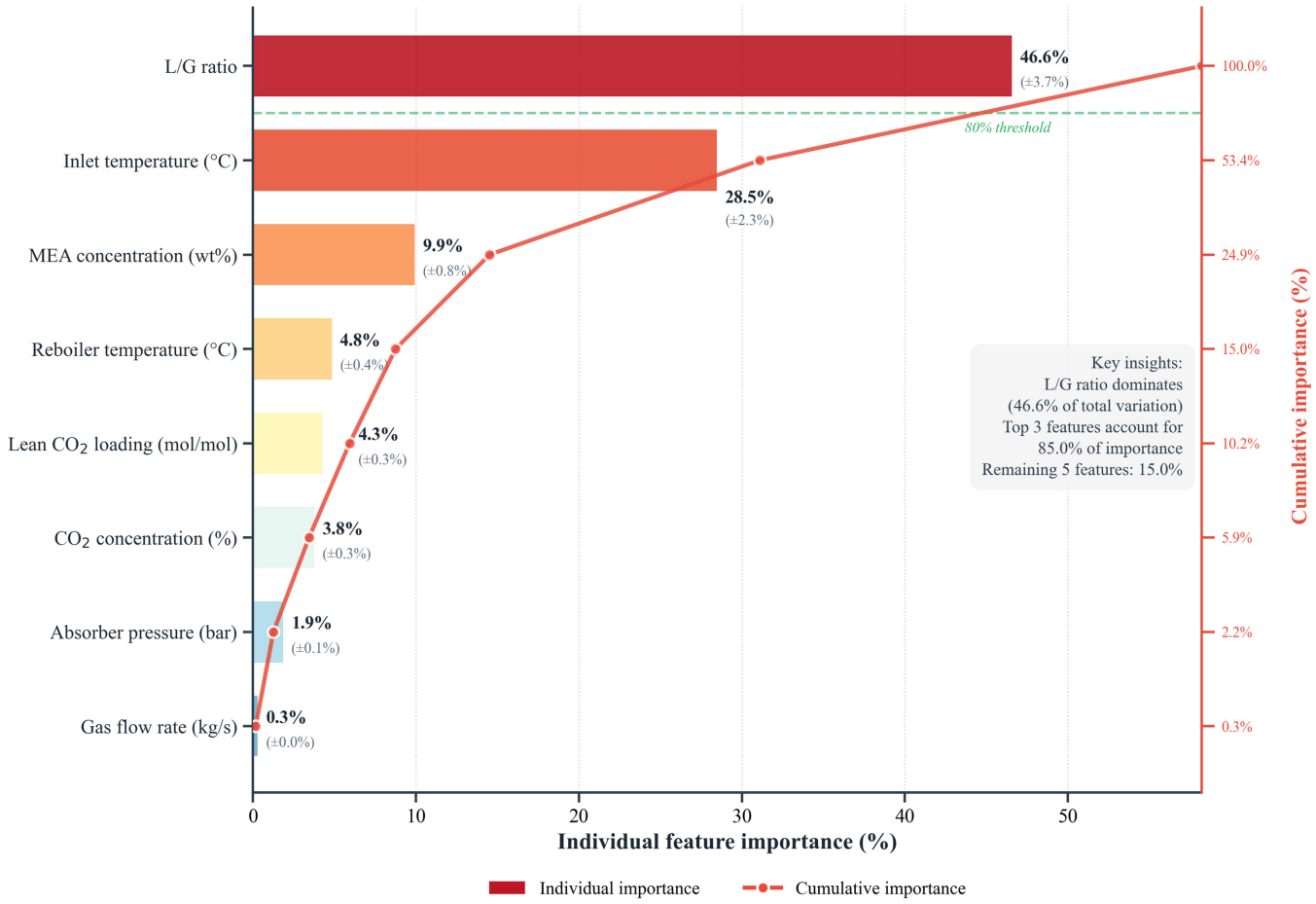

Figure 4. Feature importance hierarchy with cumulative contribution analysis.

Figure 4 quantifies the relative contribution of each operational parameter to XGBoost predictions through dual visualization, where horizontal bars represent individual feature importance percentages (x axis) and the cumulative importance curve (red line with circles) traces aggregate explanatory power building from the lowest-ranked feature at the bottom toward the highest-ranked feature at the top. The left y axis lists features ranked by importance, while the right y axis shows the corresponding cumulative importance value at each feature row, representing that feature combined with all lower-ranked features below it. Liquid-to-gas ratio emerges as the dominant predictor with 46.6% individual importance, indicated by the longest dark red bar and uncertainty bounds of ±3.7% derived

from bootstrap resampling across 50 alternative training datasets. The cumulative curve shows this single parameter, when combined with all other features, brings the total to 100.0%, as shown on the right axis at the L/G ratio row.

Inlet temperature contributes 28.5% importance (±2.3%) as the second-ranked feature, with L/G ratio and inlet temperature jointly accounting for 46.6% + 28.5% = 75.1% of total predictive power. MEA concentration adds 9.9% (±0.8%), bringing the combined contribution of the top three features to 85.0% of total importance. The annotation box confirms these key insights: L/G ratio dominates (46.6% of total variation), the top three features account for 85.0% of importance, and the remaining five features contribute only 15.0%.

The remaining five features show progressively diminishing contributions: reboiler temperature at 4.8% (±0.4%), lean CO_2 loading at 4.3% (±0.3%), CO_2 concentration at 3.8% (±0.3%), absorber pressure at 1.9% (±0.1%), and gas flow rate at 0.3% (±0.0%). These appear as progressively shorter bars with lighter colors, while the cumulative curve flattens in the lower portion of the chart as it approaches the 5.9% and 2.2% cumulative values for the three lowest-ranked features. The horizontal dashed green line at the 80% threshold intersects between the first- and second-ranked features, providing visual confirmation that the dominant L/G ratio alone (46.6%) falls below the 80% threshold, while the addition of inlet temperature brings the top two combined contribution to 75.1%, and the inclusion of MEA concentration crosses the threshold at 85.0%, validating that three parameters suffice to explain the majority of system behavior for simplified control architectures.

Uncertainty bounds in parentheses quantify bootstrap variability, revealing tight confidence intervals for high-importance features and wider relative uncertainty for minor contributors. The color gradient transitions from deep red for L/G ratio through orange and yellow to pale blue for gas flow rate, providing intuitive visual encoding where warmer colors signify higher importance. The dramatic slope change in the cumulative curve after the MEA concentration row validates dimensionality reduction strategies where operational optimization can focus on L/G ratio, inlet temperature, and MEA concentration while treating remaining parameters as secondary adjustments.

The visual hierarchy established in Figures 3 and 4 reveals liquid-to-gas ratio as the dominant predictor while demonstrating that three features collectively account for 85.0% of model importance. Table 4 provides comprehensive numerical rankings, with physical interpretation enabling systematic prioritization of operational control variables.

Table 4 consolidates SHAP feature importance rankings with physical interpretations linking statistical attributions to underlying transport phenomena. The hierarchy directly aligns with classical absorption theory, where mass transfer drivers dominate over equilibrium modifiers. Liquid-to-gas ratio governs liquid film renewal frequency and interfacial contact time, while inlet temperature controls thermodynamic boundaries through Henry's law temperature dependence. MEA concentration creates competing optimization objectives between enhanced absorption kinetics and increased regeneration penalties. This physical interpretation alignment validates model transparency and provides actionable guidance for industrial control architectures, where high-importance features require tight regulation while low-importance parameters tolerate wider operational bands.

Table 4. SHAP feature importance and physical interpretation.

Rank	Feature	Importance (%)	Cumulative (%)	Physical Interpretation
1	Liquid-to-gas ratio	46.6	46.6	Primary mass transfer driver
2	Inlet temperature	28.5	75.1	Thermodynamic limitation boundary
3	MEA concentration	9.9	85.0	Solvent absorption capacity
4	Reboiler temperature	4.8	89.8	Regeneration energy requirement
5	Lean CO_2 loading	4.3	94.1	Chemical equilibrium driving force

Table 4. *Cont.*

Rank	Feature	Importance (%)	Cumulative (%)	Physical Interpretation
6	CO_2 concentration	3.8	97.9	Reaction kinetics influence
7	Absorber pressure	1.9	99.8	Phase equilibrium effect
8	Gas flow rate	0.3	100.1 *	Contact time modulator

Notes: * Rounding to 100.0% in practice. The cumulative (%) column aggregates feature importance sequentially from the highest-ranked to the lowest-ranked feature, following the conventional top-down ranking convention for tabular presentation. This differs directionally from the right *y* axis of Figure 4, which accumulates from the least important feature upward to visualize the shape of the importance distribution curve. Liquid-to-gas ratio dominates feature importance at 46.6%, representing nearly half of total predictive influence, confirming its role as primary operational control variable as quantified through Equation (16), with percentage contributions normalized via Equation (17). The top three features include liquid-to-gas ratio, inlet temperature, and MEA concentration, collectively accounting for 85.0% cumulative importance, establishing clear hierarchy for process optimization priorities. The two temperature-related control parameters, inlet temperature (28.5%) and reboiler temperature (4.8%), together contribute 33.3% of total importance, highlighting the critical role of thermal management in both absorption driving force and regeneration energy requirements. Lower-ranked features, including absorber pressure at 1.9% and gas flow rate at 0.3%, exhibit minimal individual impact yet remain necessary for comprehensive system characterization. Packing height and solvent circulation rate were not included in the SHAP decomposition as preliminary sensitivity screening indicated negligible influence on capture efficiency predictions, and their exclusion does not affect the reported importance hierarchy. Physical interpretation alignment validates model transparency where mass transfer drivers dominate over equilibrium modifiers, consistent with absorption column design principles established in Section 2.2.

3.3. Parametric Response Analysis and Optimal Operating Regions

SHAP importance analysis quantified through Equation (16) establishes which features dominate prediction variance, while partial dependence analysis reveals how efficiency responds functionally to parametric changes. Response surfaces isolate individual feature effects by computing conditional expectations through averaging over marginal distributions, enabling detection of nonlinear optimization curves and saturation phenomena. Figure 5 characterizes response surfaces for the top three features from Table 4 using 95% confidence bands derived from 50 bootstrap samples calculated via Equation (22).

Panel A demonstrates the liquid-to-gas ratio response surface exhibiting a strong positive monotonic relationship where efficiency increases from approximately 76% at L/G ratio 2.0 to a plateau at 97.8% achieved at the optimal value of 6.29 (red star). The mean effect curve (dark blue line with circles) shows a steep initial slope between 2.1 and 4.5, indicating high marginal returns from solvent circulation increases, followed by gradual saturation beyond 5.0 where the curve flattens approaching asymptotic maximum efficiency. The 95% confidence interval (light blue shaded region) widens progressively at higher L/G ratios, reflecting greater prediction uncertainty in sparsely sampled parameter regions, while raw data points (light blue scatter with error bars showing ± 1 SD) populate the feasible operating space. The optimal operating region (light green shaded area) spans L/G ratios from approximately 4.5 to 6.5 where efficiency exceeds 97%.

Panel B reveals the inlet temperature response exhibiting characteristic inverted U-shaped curve where efficiency peaks at 95.8% at the optimal temperature of 40.33 °C (purple star), then declines on both sides. Below 40 °C, the mean effect curve (dark red line with circles) shows modest efficiency reduction to approximately 95% at 32 °C, while above 45 °C efficiency drops more dramatically, reaching 88% at 50 °C and declining to approximately 78% at 53 °C, demonstrating severe thermodynamic penalties from elevated temperatures. The 95% confidence interval (light orange shaded region) remains relatively narrow across the temperature range, indicating robust predictions. The optimal operating region (light orange shaded area) spans approximately 32 °C to 46 °C, providing a 14-degree operational window for near-maximum efficiency. This nonlinear response validates the negative SHAP values observed for high temperatures in Figure 3.

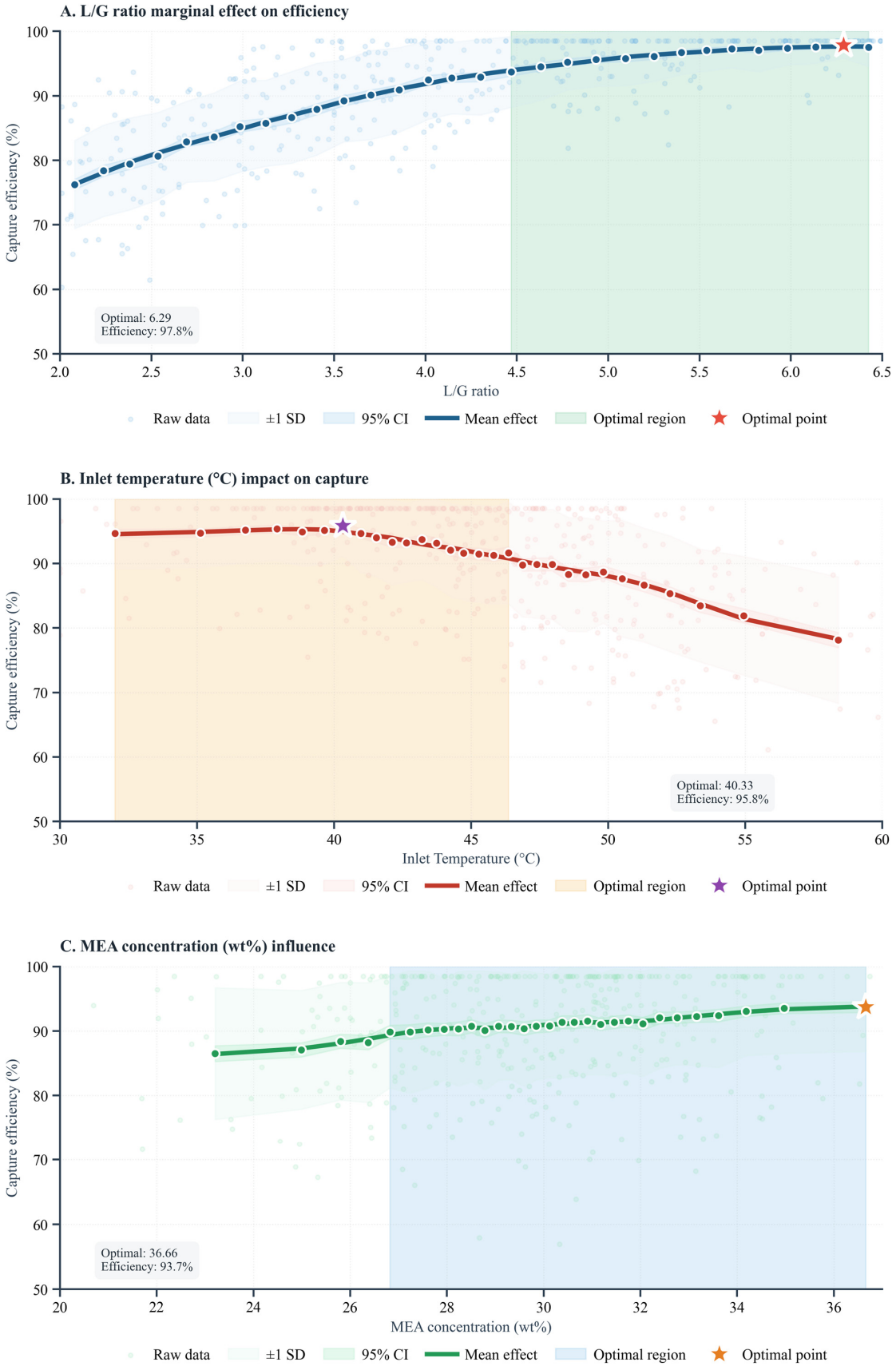

Figure 5. Partial dependence analysis revealing nonlinear response surfaces and optimal operating windows for top three features.

Panel C displays MEA concentration response, showing a moderate positive relationship where efficiency increases from approximately 86% at 23 wt% to 93.7% at the optimal concentration of 36.66 wt% (orange star). The mean effect curve (dark green line with circles) exhibits a nearly linear trend with a modest slope, indicating lower sensitivity compared to L/G ratio and inlet temperature. The 95% confidence interval (light blue shaded region) expands at concentration extremes, particularly above 35 wt%, suggesting greater prediction uncertainty near solvent degradation limits. The optimal operating region (light blue shaded area) spans broadly from approximately 27 wt% to 36.3 wt%, encompassing conventional 30 wt% baseline formulations and extending to concentrated advanced solvents. Raw data scatter demonstrates high density across the concentration range with relatively uniform distribution.

Univariate partial dependence analysis in Figure 5 reveals individual parameter response surfaces while holding other features at average values, effectively isolating marginal effects. However, real process behavior involves simultaneous variation in multiple parameters, where interaction effects may produce synergistic or antagonistic relationships not captured through one-dimensional analysis. Figure 6 extends the analysis to bivariate surfaces for the three most important feature pairs identified through Equation (18), revealing optimal operating regions that balance competing drivers. Interaction strength quantified through Equation (19) enables the detection of coupling effects, while two-way interaction sensitivity (Equation (23)) quantifies synergistic deviations from additive behavior.

Panel A displays liquid-to-gas ratio and inlet temperature interaction, exhibiting antagonistic relationship where optimal efficiency of 98.6% occurs at L/G ratio 6.44 combined with inlet temperature 40.00 °C (red star). The contour map reveals steep efficiency gradient with deep blue coloring in the upper left quadrant, indicating poor performance below 60% efficiency at combinations of low L/G ratios (2.0 to 2.5) and high temperatures (60 °C to 65 °C), demonstrating severe penalties when unfavorable conditions coincide. Conversely, the lower right quadrant exhibits dark red coloring, representing efficiency exceeding 90% across broad operating regions where elevated L/G ratios (3.6 to 7.0) combine with moderate inlet temperatures (30 °C to 50 °C), establishing a favorable operating window. The contour lines progress from widely spaced intervals in the high-efficiency region to tightly packed lines in the transition zone between 70% and 84% efficiency, signifying rapid performance degradation as conditions deviate from optimal. Scattered data points populate the feasible parameter space with higher density in the central region. The interaction surface demonstrates nonlinear curvature, where optimal temperature shifts from approximately 45 °C at low L/G ratios to 38 °C at high ratios, indicating that thermal management requirements depend on solvent circulation intensity.

Panel B illustrates liquid-to-gas ratio and MEA concentration interaction, displaying a strong synergistic relationship where an optimal efficiency of 99.1% is achieved at L/G ratio 4.85 combined with MEA concentration 38.34 wt% (red star). The contour map exhibits diagonal efficiency gradient progressing from deep purple in the lower left corner representing approximately 70% efficiency at low L/G ratios (2.0 to 2.2) and low concentrations (20 wt%) to bright yellow in the upper right corner indicating 98% efficiency at high L/G ratios (5.0 to 6.5) and high concentrations (37 to 40 wt%). This diagonal orientation reveals that efficiency improvements require coordinated increases in both parameters rather than optimization of either variable independently. The contour lines maintain relatively uniform spacing across the parameter space, suggesting approximately linear interaction without dramatic threshold effects, contrasting with the sharp transition zones observed in Panel A. The broad yellow region spanning L/G ratios from 4.7 to 6.5 and MEA concentrations from 31 to 40 wt% defines a generous optimal operating window

where efficiency consistently exceeds 95%, providing substantial operational flexibility. The synergistic interaction implies that investments in increased solvent circulation yield greater returns when paired with concentrated MEA formulations.

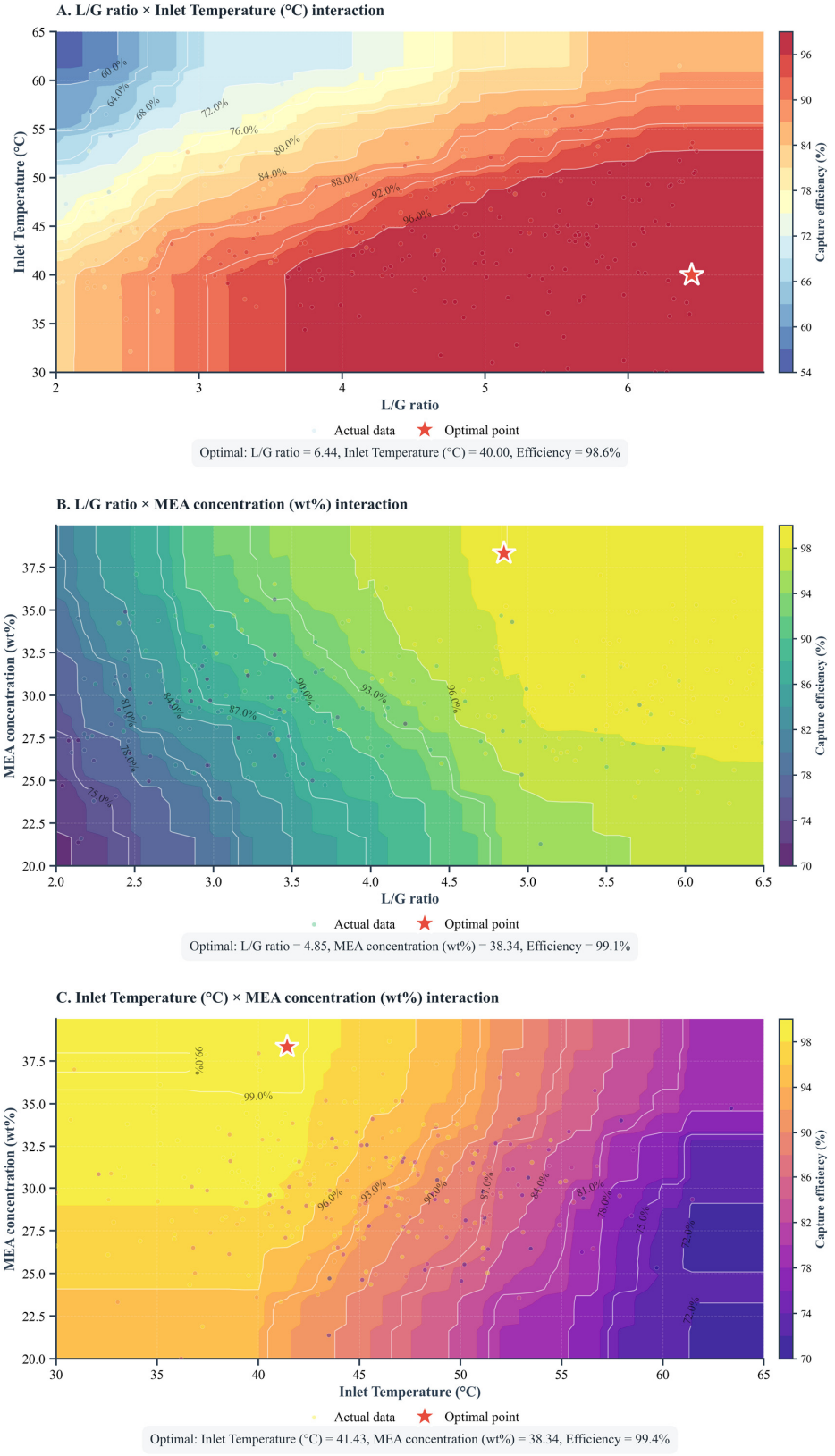

Figure 6. Two-way feature interaction contour maps revealing synergistic optimization opportunities.

Panel C depicts inlet temperature and MEA concentration interaction, exhibiting a complex nonlinear relationship where an optimal efficiency of 99.4% occurs at inlet temperature 41.43 °C combined with MEA concentration 38.34 wt% (red star). The contour map displays a distinctive asymmetric pattern with a bright yellow region concentrated in the upper left quadrant spanning temperatures 30 °C to 45 °C and concentrations 35 to 40 wt%, representing the high-efficiency zone above 98%. Moving rightward toward higher temperatures, the contours transition rapidly through orange and pink to deep purple, indicating efficiency collapse below 74% at temperatures exceeding 60 °C regardless of MEA concentration, confirming the dominant thermodynamic penalty imposed by elevated temperature. The left portion of the map maintains moderate efficiency above 82%, even at low concentrations (20 to 25 wt%) when temperatures remain below 45 °C, demonstrating that thermal control can partially compensate for dilute solvent formulations. Contour line curvature reveals that optimal temperature decreases slightly from approximately 43 °C to 40 °C as concentration increases from 25 to 38 wt%, suggesting weak interaction where concentrated solvents prefer marginally cooler operation. The narrow optimal region compared to Panel B indicates tighter operational tolerances requiring precise coordinated control of both thermal management and solvent concentration.

The bivariate interaction surfaces presented in Figure 6 identify optimal operating points for parameter pairs with a maximum predicted efficiency of 99.4% achieved through temperature and concentration coordination. However, comprehensive operational guidelines require simultaneous optimization of all parameters. Table 5 provides systematic operating targets derived from bi-objective optimization (Equations (24) and (25)), satisfying Pareto-optimality criteria (Equation (29)) for efficiency levels from 85% to 98%, derived from Pareto-optimal solutions that balance efficiency maximization against regeneration duty minimization under operational constraints. Parameters not included in Table 5, specifically lean CO_2 loading, absorber pressure, packing height, and gas flow rate, are held at their Table 1 mean values of 0.268 mol/mol, 1.13 bar, 15.2 m, and 150.0 kg/s, respectively, consistent with their combined SHAP importance contribution of only 10.2%, confirmed in Table 4.

The optimization results reveal trade-offs governing process economics. Achieving 85% capture requires 3.1 MJ/kg CO_2, while 98% capture demands 4.3 MJ/kg CO_2, representing a 39% energy penalty for 13 percentage points of additional efficiency. L/G ratio emerges as the primary control variable, scaling nonlinearly from 3.2 to 6.3, with strongly diminishing returns, where the first 5 percentage point gain from 85% to 90% requires a 31% L/G increase from 3.2 to 4.2, whereas the final 3 percentage point gain from 95% to 98% requires only a 14.5% L/G increase from 5.5 to 6.3 while carrying disproportionately higher energy and cost penalties, as reflected in the SRD escalation from 3.9 to 4.3 MJ/kg CO_2 and cost index jump from 1.42 to 1.68.

Inlet temperature specifications follow a non-monotonic optimum rather than a simple monotonic constraint, where the highest-efficiency tier of 98% requires temperatures within the narrow window of 38 °C to 43 °C, consistent with the global optimum of 99.4% identified at 41.43 °C in Figure 6 Panel C. Operating outside this window in either direction incurs penalties through reduced reaction kinetics at lower temperatures or thermodynamic solubility limitations at elevated temperatures, as validated by the response surfaces in Figure 5 Panel B. At lower efficiency tiers, wider inlet temperature tolerances of up to 50 °C for the 85% target reflect greater operational flexibility where thermodynamic constraints are less binding.

The relative cost index is a dimensionless operational cost proxy computed as the normalized weighted sum of specific regeneration duty (70% weight) and L/G ratio contribution (30% weight), normalized to unity at the 85% baseline, capturing the dominant cost

drivers identified through SHAP analysis without requiring full techno-economic modeling. This index shows that 90% capture incurs a 15% premium over baseline, while 95% and 98% targets require 42% and 68% cost increases, respectively. These guidelines enable operating point selection aligned with project constraints. Stringent emission regulations may justify the 68% premium for maximum capture, while cost-sensitive applications might optimize around 90% efficiency for carbon reduction at only a 15% incremental cost. MEA concentration windows shifting from 25 to 35 wt% at the 85% tier to 32–40 wt% at the 98% tier reflect the synergistic interaction between solvent concentration and L/G ratio confirmed in Figure 6 Panel B, where higher solvent concentrations at elevated L/G ratios amplify absorption driving force through combined enhancement of interfacial area and reaction kinetics.

Table 5. Optimized operational guidelines and performance targets.

Target Efficiency (%)	L/G Ratio (Optimal)	Inlet Temp. (°C)	MEA Conc. (wt%)	Reboiler Temp. (°C)	Expected SRD (MJ/kg CO_2)	Relative Cost Index
≥85	3.2	≤50	25–35	110	3.1	1.00
≥90	4.2	≤45	28–35	115	3.4	1.15
≥95	5.5	≤40	30–38	120	3.9	1.42
≥98	6.3	38–43	32–40	125	4.3	1.68

Notes: Progressive efficiency targets require systematic L/G ratio increases from 3.2 to 6.3, exhibiting diminishing returns where the final 3% efficiency gain demands a 14.5% increase in solvent circulation. Inlet temperature follows a non-monotonic optimum, where the highest-efficiency targets require temperatures within the narrow window of 38 °C to 43 °C, consistent with the global optimum of 99.4% identified at 41.43 °C in Figure 6 Panel C; operating below 35 °C or above 50 °C at high-efficiency targets incurs penalties through reduced reaction kinetics or thermodynamic solubility limitations, respectively, as validated by the inverted U-shaped response surface in Figure 5 Panel B. Specific regeneration duty escalates from 3.1 to 4.3 MJ/kg CO_2, representing a 39% energy penalty when pursuing 98% capture versus 85% baseline. Relative cost index indicates that 95% efficiency incurs 42% higher operational costs compared to 85% baseline, while 98% target requires a 68% cost premium. MEA concentration windows shift toward higher ranges at elevated efficiency targets (25–35 wt% at 85% to 32–40 wt% at 98%), consistent with synergistic relationships in Figure 6 Panel B. Guidelines are derived from Pareto-optimal solutions (Equation (29)), satisfying the efficiency constraint (Equation (27)) and thermal degradation limit (Equation (28)) while minimizing the weighted objective function (Equation (32)). Solution diversity is maintained through the crowding distance metric (Equation (30)). Trade-off slopes (Equation (31)) quantify marginal energy costs per efficiency gain, with bi-objective problem formulation defined in Equations (24) and (25). Relative cost index is a dimensionless operational cost proxy computed as the normalized weighted sum of specific regeneration duty contribution (70% weight, reflecting dominant energy operating cost) and L/G ratio contribution (30% weight, reflecting solvent pump and circulation capital cost), normalized to unity at the 85% baseline efficiency operating point; this proxy captures the dominant cost drivers identified through SHAP analysis in Table 4 without requiring full techno-economic modeling. Parameters not listed in Table 5, including lean CO_2 loading (0.268 mol/mol), absorber pressure (1.13 bar), and gas flow rate (150.0 kg/s), are held at their mean values from Table 1 across all efficiency tiers, consistent with SHAP importance analysis in Table 4, confirming their combined contribution of only 10.2% to total predictive variance; these parameters represent secondary adjustments that can be fixed at nominal operating conditions without materially affecting the efficiency targets specified. Packing height (mean 15.2 m) and solvent circulation rate (mean 245.0 m^3/h) are similarly fixed at their Table 1 mean values, as preliminary sensitivity screening indicated negligible influence on capture efficiency predictions.

While Figure 6 and Table 5 establish optimal operating points through deterministic optimization, operational robustness requires understanding of how parameter variations propagate through the system. Figure 7 presents global sensitivity analysis quantifying parametric influence on capture efficiency through systematic perturbations calculated via local sensitivity coefficients defined in Equation (20). Bootstrap uncertainty quantification employs confidence intervals computed through Equations (21) and (22), while multi-level perturbations spanning ±30% variations enable assessment of nonlinear response patterns and interaction effects quantified through Equation (23). This analysis provides insights for robust process control and uncertainty management.

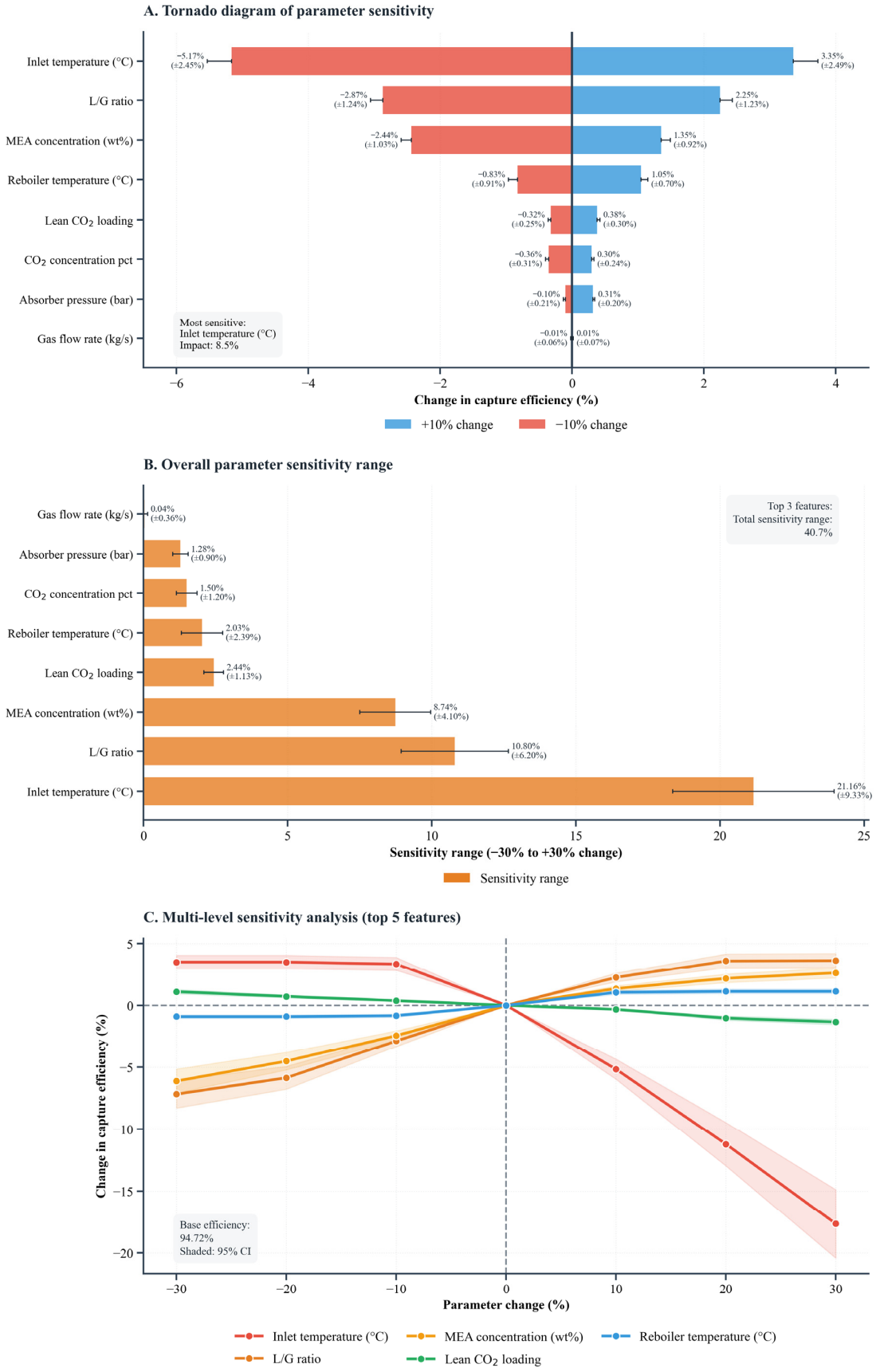

Figure 7. Global sensitivity analysis of process parameters influencing CO_2 capture efficiency.

Figure 7 presents comprehensive sensitivity analysis quantifying the influence of eight key process parameters on CO_2 capture efficiency within an amine-based absorption

system. Panel A employs a tornado diagram illustrating changes in capture efficiency resulting from ±10% deviations from baseline parameter values, with horizontal axes denoting percentage change in efficiency and vertical axes listing parameters ordered by total impact magnitude. Inlet temperature exhibits the greatest sensitivity, where a +10% increase improves efficiency by 3.35 percentage points while a −10% reduction decreases it by 5.17 percentage points, demonstrating inverse response behavior. The liquid-to-gas ratio follows with a 2.25 percentage point gain and a 2.87 percentage point loss for +10% and −10% variations, respectively. Monoethanolamine (MEA) concentration produces a 1.35 percentage point gain and a 2.44 percentage point loss, while reboiler temperature shows a 1.05 percentage point gain and a 0.83 percentage point loss. At the opposite end of the sensitivity spectrum, gas flow rate exhibits minimal influence, with only a 0.01 percentage point change for ±10% perturbations, while absorber pressure produces negligible a 0.10 percentage point variation, confirming these parameters exert limited control over capture performance. Error bars derived from bootstrap iterations quantify prediction uncertainty across sampled conditions.

Panel B complements this analysis through horizontal bar charts, depicting overall sensitivity ranges computed as differences between maximum and minimum efficiency values across −30% to +30% parameter perturbations. Inlet temperature spans the widest interval at 21.16 percentage points, followed by L/G ratio (10.80 percentage points) and MEA concentration (8.74 percentage points), with the three most influential parameters collectively accounting for 40.7 percentage points of total sensitivity range. In contrast, gas flow rate displays the narrowest sensitivity range at only 0.04 percentage points, followed by absorber pressure at 1.28 percentage points and CO_2 concentration at 1.50 percentage points, collectively representing less than 3 percentage points of total variation. Error bars reflect the combined uncertainty from extreme parameter variations, underscoring methodological robustness.

Panel C extends the investigation through multi-level sensitivity curves tracing efficiency responses across continuous −30% to +30% parameter shifts for the five most impactful variables, where horizontal axes represent percentage changes and vertical axes indicate corresponding efficiency deviations from baseline 94.72%. Inlet temperature demonstrates strong negative response progressing from +4.3 percentage points at −30% to −17.4 percentage points at +30%, confirming dominant thermodynamic constraints. L/G ratio exhibits an upward trajectory ranging from −7.0 percentage points at −30% to +3.0 percentage points at +30%, demonstrating positive correlation with efficiency. MEA concentration follows a similar pattern, progressing from −7.0 percentage points at −30% to +4.8 percentage points at +30%, reflecting enhanced absorption capacity at higher concentrations. Reboiler temperature shows a modest positive response, spanning from −0.1 percentage points at −30% to +1.5 percentage points at +30%, indicating relatively minor influence compared to dominant parameters. Lean CO_2 loading maintains slight positive sensitivity with response magnitude confined within −0.8 to +0.2 percentage points across the entire perturbation range, representing the weakest influence among the featured parameters. Shaded 95% confidence intervals quantify propagation of modeling uncertainty through response surfaces.

The analysis reveals that absorption performance depends on mass transfer rates controlled by L/G ratio and MEA concentration, while temperature effects exhibit strong inverse sensitivity explained by CO_2 solubility thermodynamics in aqueous amine solutions. The systematic variation across perturbation magnitudes confirms model nonlinearity, with response curves deviating from proportional scaling beyond ±10% deviations.

3.4. External Validation Against CASTOR Pilot Data

External validation used the CASTOR monoethanolamine pilot data described in Section 2.7, where flue-gas and solvent conditions from ten operating points were mapped onto the ten-dimensional input vector of the XGBoost surrogate. For each point, the measured flue-gas and solvent conditions defined the boundary conditions for the mechanistic model and for the surrogate, while the reported CO_2 removal rate and regeneration energy served as experimental references. Pilot plant operating points and prediction errors are documented in Supplementary Materials, File S8: 08_CASTOR_MEA_Validation.csv.

Percentage deviations were computed using Equations (33) to (35), with capture efficiency and specific regeneration duty considered separately. Across the ten CASTOR points, the XGBoost model overpredicted capture efficiency with $MAPD_{\eta} = 65.6\%$ and $RMSPD_{\eta} = 66.6\%$, and it underpredicted specific regeneration duty with $MAPD_{SRD} = 16.7\%$ and $RMSPD_{SRD} = 17.7\%$. Pearson correlation coefficients were $r = -0.23$ for efficiency and $r = 0.41$ for specific regeneration duty, which indicates limited linear agreement between the surrogate and the pilot data at this small number of operating points.

4. Discussion

The integration of machine learning surrogates with first-principles simulation frameworks addresses a fundamental challenge in carbon capture system optimization, where rigorous rate-based models impose latency constraints that often preclude comprehensive parameter space exploration and real-time process control [54,55]. This investigation reveals that XGBoost regression models trained on 10,000 simulation-derived operating conditions achieve R^2 of 0.9701 with a prediction latency of merely 1.5 milliseconds, as documented in Table 3, enabling design space evaluations that would require prohibitive computational resources using conventional rate-based simulation platforms. Neural network architectures achieve superior predictive accuracy at R^2 of 0.9729 yet demand 45.3 s training duration compared to 12.4 s for XGBoost, making gradient boosting methods preferable for iterative model updating scenarios encountered in adaptive control implementations.

This computational architecture enables digital twin deployment where real-time operational data feeds SHAP analysis modules that continuously update feature importance rankings and optimal operating targets, generating adaptive control signals responding to fluctuating feed compositions or energy prices within sub-second latencies, infeasible with rigorous simulation. The 38 MB memory footprint enables deployment on resource-constrained embedded controllers suitable for distributed process control architectures, eliminating dependencies on centralized computation infrastructure that introduce communication latencies incompatible with closed-loop control timing requirements. The throughput of 667 predictions per second achieved by XGBoost translates to a capacity for evaluating 40,000 design alternatives per minute, enabling Monte Carlo uncertainty quantification across million-case ensembles within practical computational budgets.

SHAP interpretability analysis reveals that liquid-to-gas ratio commands 46.6% of total model importance, and the resulting feature hierarchy is consistent with established transport phenomena principles, highlighting that SHAP-based attributions can align data-driven models with mechanistic understanding [56]. The quantitative precision enabled through Shapley value decomposition transforms qualitative engineering intuition into numerical hierarchies amenable to formal control system design, instrumentation prioritization, and process economics analysis. Inlet temperature accounts for 28.5% of the variable importance and exhibits a characteristic antagonistic relationship where elevated temperatures impose thermodynamic penalties through reduced CO_2 solubility governed by Henry's law while simultaneously enhancing chemical reaction rates via Arrhenius kinetics. This duality creates optimization complexity absent in isothermal processes,

requiring sophisticated control strategies that balance competing thermodynamic and kinetic drivers across operating conditions.

The cumulative importance of 85.0% achieved by only three parameters suggests viable reduced-order control architectures focusing on liquid-to-gas ratio, inlet temperature, and MEA concentration while treating remaining variables as secondary adjustments or fixed design specifications. This dimensionality reduction addresses practical implementation constraints where industrial distributed control systems impose limits on manipulated variable counts and control loop complexity, necessitating parsimonious strategies that capture dominant system behavior without exhaustive instrumentation. The hierarchical structure identified through SHAP analysis provides objective criteria for systematically eliminating low-importance parameters from active control consideration, enabling streamlined automation architectures that reduce capital costs while maintaining near-optimal performance.

Partial dependence analysis quantifies nonlinear response surfaces, revealing that efficiency increases from 76% at L/G ratio 2.0 to plateau at 97.8% at optimal value 6.29, as shown in Figure 5 Panel A, demonstrating saturation behavior characteristic of mass transfer-limited processes approaching thermodynamic equilibrium ceilings. The steep initial gradient below L/G 4.0 yielding approximately a 5 percentage points efficiency gain per unit ratio increase reflects liquid side resistance reduction through enhanced interfacial contact, while subsequent flattening beyond 5.0 indicates diminishing returns where additional solvent circulation provides marginal benefit as gas-side resistance and equilibrium limitations become dominant. This transition demarcates economically optimal operating regions where capital investments in pump capacity and piping infrastructure deliver substantial returns below the transition point but encounter diminishing productivity beyond saturation thresholds. A temperature response exhibiting inverted U-shaped optimization peaking at 40.33 °C with maximum efficiency of 95.8%, as depicted in Figure 5 Panel B, quantifies the compromise between solubility-driven thermodynamic penalties at elevated temperatures and kinetic sluggishness at reduced temperatures, establishing narrow optimal windows requiring precise thermal management.

Bivariate interaction surfaces extend univariate analysis, revealing synergistic relationships between liquid-to-gas ratio and MEA concentration, where an optimal efficiency of 99.1% at an L/G ratio 4.85 combined with MEA concentration 38.34 wt% requires coordinated parameter increases rather than independent optimization, as demonstrated in Figure 6 Panel B. The diagonal contour progression demonstrates that circulation intensity and solvent concentration amplify each other's effectiveness, likely through mechanisms where concentrated amine solutions more effectively exploit enhanced interfacial contact provided by elevated liquid flow rates. This synergy informs capital budgeting decisions, suggesting integrated investments that simultaneously upgrade pumping capacity and solvent concentration rather than staged sequential improvements that fail to capture interaction benefits. Conversely, antagonistic relationships between liquid-to-gas ratio and inlet temperature shown in Figure 6 Panel A warn against operating regimes combining insufficient circulation with elevated temperatures, where compounding thermodynamic and mass transfer deficiencies produce efficiency collapse below 60%. The maximum predicted efficiency of 99.4% achieved at inlet temperature 41.43 °C combined with MEA concentration 38.34 wt% in Figure 6 Panel C establishes the global optimum across all parameter combinations examined.

Multi-objective optimization generates operational guidelines spanning efficiency targets from 85% to 98% that translate Pareto-optimal solutions into actionable engineering specifications, following established practices for NSGA-II based process optimization [44,57]. The progressive escalation from L/G ratio 3.2 at 85% efficiency to 6.3 at 98%

efficiency quantifies nonlinear resource scaling, where final incremental gains demand disproportionate investments, informing cost–benefit analyses under heterogeneous regulatory and economic scenarios. Specific regeneration duty escalation from 3.1 to 4.3 MJ/kg CO_2, representing a 39% energy penalty, establishes the fundamental efficiency–energy trade-off constraining system economics, enabling transparent multi-criteria decision frameworks where stakeholders explicitly weigh environmental objectives against operational expenditures. The relative cost index progression quantified in Table 5 provides decision support tools where emission-intensive facilities facing stringent regulations or high carbon prices identify optimal operating points at elevated efficiency targets, while cost-sensitive applications optimize toward lower efficiency thresholds, achieving substantial carbon reduction at modest incremental costs. Specifically, 90% capture incurs only a 15% premium over 85% baseline, while 95% and 98% targets require 42% and 68% cost increases, respectively.

Sensitivity analysis quantified in Figure 7 reveals that inlet temperature exhibits the widest parametric influence, with a sensitivity range spanning 21.16 percentage points across the −30% to +30% perturbation domain, followed by an L/G ratio at 10.80 percentage points and an MEA concentration at 8.74 percentage points. The three most influential parameters collectively account for 40.7 percentage points of the total sensitivity range, confirming their dominant role in process control. Panel C multi-level curves demonstrate that inlet temperature sensitivity progresses from +4.3 percentage points at −30% to −17.4 percentage points at +30% relative to a baseline efficiency of 94.72%, while the L/G ratio spans from −7.0 to +3.0 percentage points and MEA concentration ranges from −7.0 to +4.8 percentage points across the same perturbation range. The systematic variation across perturbation magnitudes confirms model nonlinearity, with response curves deviating from proportional scaling beyond ±10% deviations.

The methodology demonstrates broader applicability extending beyond carbon capture to chemical process systems where computational expense traditionally constrains design optimization. The workflow combining Latin hypercube sampling for efficient parameter space coverage, ensemble machine learning for accurate surrogate construction, SHAP analysis for mechanistic interpretability, and evolutionary algorithms for multi-objective optimization establishes a generalizable framework transferable across distillation, reaction engineering, crystallization, and separation process contexts. This transferability suggests that process systems engineering stands at an inflection point where machine learning tools transition from niche research applications to standard industrial practice, fundamentally altering how chemical plants are designed, operated, and optimized.

The explicit knowledge extraction enabled through SHAP decomposition addresses longstanding industrial skepticism toward black box models by providing instance-level explanations, revealing why specific operating conditions produce efficiency predictions. TreeExplainer algorithm computing exact Shapley values with polynomial complexity makes these explanations computationally feasible for industrial-scale applications, where real-time transparency requirements preclude expensive permutation sampling approaches. This interpretability facilitates regulatory approval processes and safety certification procedures, where authorities demand mechanistic justification for automated control decisions affecting critical infrastructure. The alignment between SHAP importance rankings and classical transport phenomena theory validates that machine learning models learn physically consistent relationships rather than spurious correlations, establishing credibility for deployment in safety-critical applications.

The convergence of five independent machine learning architectures toward R^2 exceeding 0.94 provides internal cross-validation that diverse algorithmic approaches consistently approximate simulation-derived thermodynamic relationships, though this agreement

with simulation outputs should be distinguished from validation against physical pilot data where MAPD of 65.6% for capture efficiency indicates that further refinement is required. This convergence reduces concerns about model-specific artifacts or overfitting to training configurations, though external pilot-scale validation remains essential before quantitative deployment. This consensus across gradient boosting, neural networks, random forests, and support vector methods suggests fundamental learnability of the absorption system input–output mapping, indicating that predictive success stems from genuine pattern recognition rather than algorithmic peculiarities. The methodology positions carbon capture optimization within broader digital transformation trends affecting chemical manufacturing, where data-driven approaches complement rather than replace physics-based engineering, creating hybrid paradigms that leverage the strengths of both frameworks while mitigating their respective limitations.

5. Conclusions

This study established an integrated digital twin-oriented framework that couples mechanistic process models with explainable machine learning to support systematic optimization of post-combustion MEA scrubbing systems. Latin hypercube sampling [58] generated 10,000 diverse operating scenarios evaluated through equilibrium stage mass transfer models incorporating electrolyte NRTL thermodynamics, zwitterion reaction mechanisms, and film theory enhancement factors [59]. XGBoost surrogate models achieved R^2 of 0.9701 with a 1.5 millisecond inference latency and 667 predictions per second throughput, while SHAP decomposition quantified feature importance establishing liquid-to-gas ratio (46.6%), inlet temperature (28.5%), and MEA concentration (9.9%) as dominant efficiency determinants. Pareto frontier identification generated operational guidelines spanning 85% to 98% capture efficiency with regeneration duties from 3.1 to 4.3 MJ/kg CO_2, quantifying the complete efficiency–energy cost trade-off landscape. Pareto-optimal operating points are provided in Supplementary Materials, File S7: 07_Pareto_Optimal_Solutions.csv.

The key methodological contribution lies in demonstrating that Shapley value attribution grounded in game-theoretic axioms of efficiency, symmetry, dummy, and additivity enables mechanistic validation of surrogate model behavior against classical transport phenomena theory. Bivariate interaction analysis identified synergistic relationships between L/G ratio and MEA concentration, achieving 99.1% efficiency through coordinated optimization, contrasting with antagonistic temperature interactions that produce efficiency collapse below 60% when unfavorable conditions coincide. The maximum predicted efficiency of 99.4% at 41.43 °C inlet temperature and 38.34 wt% MEA concentration establish global optima across the ten-dimensional parameter space. Sensitivity analysis revealed inlet temperature dominance with a 21.16 percentage point range, validating importance rankings derived independently through SHAP analysis and confirming methodological consistency.

Practical limitations remain, because the training database relies entirely on simulations and external validation is restricted to ten CASTOR pilot operating points analyzed through the metrics in Section 3.4. The XGBoost surrogate exhibited mean absolute percentage deviations of 65.6% for capture efficiency and 16.7% for specific regeneration duty, indicating that the present model is not yet calibrated for quantitative reproduction of pilot measurements even though it captures consistent trends across the simulated design space. The SHAP-based explanations are additionally subject to systematic failure modes identified by Bove et al. [60]: At the model level, SHAP attributions are only as faithful as the XGBoost surrogate and may therefore reflect simulation artifacts rather than true physical mechanisms given the CASTOR validation gap; at the user level, plant operators may misinterpret SHAP values as causal operational recommendations whereas SHAP

provides correlational attribution within the training distribution only; and at the deployment level, operating conditions outside the training domain such as degraded solvent states or atypical flue-gas compositions may produce unreliable attributions. The bootstrap robustness analysis across 50 training resamples partially addresses model-level instability by confirming stable feature rankings across resampled datasets, but these broader XAI failure modes remain active limitations requiring further investigation.

Future work should expand the validation set using additional pilot campaigns and refine thermodynamic correlations so that simulation-based data generation aligns more closely with experimental observations across wider operating windows. The current parameter ranges encompass conventional MEA absorption envelopes but exclude emerging hybrid configurations incorporating membrane contactors or blended solvent formulations that may offer performance improvements. Bootstrap uncertainty intervals quantify surrogate prediction variability alone rather than combined simulation and surrogate uncertainty, which may underestimate total prediction error when the underlying physics model contains approximations.

This research demonstrates that knowledge extraction from process simulations through interpretable machine learning transforms carbon capture optimization from computationally constrained empiricism to systematic data-driven discovery, revealing parameter interactions and operating strategies inaccessible through traditional approaches. The methodology positions artificial intelligence as a discovery tool augmenting mechanistic insight rather than replacing engineering understanding, ultimately accelerating the deployment of emission reduction technologies for climate change mitigation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/make8040087/s1, The primary simulation and analysis files are: 01_Full_Simulation_Dataset.csv (physics-based operating conditions and resulting performance indicators); 02_ML_Model_Comparison.csv (benchmark metrics for the surrogate models); 03_SHAP_Global_Feature_Importance.csv (global feature importance values from SHAP analysis); 04_SHAP_Detailed_Values.csv (instance-level SHAP values and prediction errors); 05_Decision_Tree_Rules.txt (interpretable decision tree rules extracted from the XGBoost surrogate); 06_Operational_Guidelines.csv (data-driven operational rules for different capture efficiency targets); 07_Pareto_Optimal_Solutions.csv (Pareto-optimal operating points for efficiency and specific regeneration duty); 08_CASTOR_MEA_Validation.csv (pilot plant operating points and prediction errors used for external validation of the surrogate); 09_Computational_Benchmark.csv (computational performance comparison between rigorous simulation and surrogate models); 10_Final_Summary_Report.csv (summary indicators for the complete workflow and validation outcomes).

Funding: This research received no external funding.

Data Availability Statement: The csv data files that support the findings of this study are provided as Supplementary Files uploaded with the manuscript to the MDPI's submission system. All datasets were generated using the Python-based simulation and machine learning pipeline described in Section 2 and used directly for the analyses reported in Sections 3 and 4. All files will be available as Supplementary Materials through the MDPI platform upon publication. Researchers may also request the datasets and the associated Python scripts from the corresponding author for further reuse or verification.

Acknowledgments: This research was supported by the "University of Debrecen Program for Scientific Publication". During preparation of this manuscript, DeepL Write (https://www.deepl.com/en/translator, 10 February 2026) and ScienceDirect AI (https://www.elsevier.com/products/sciencedirect/sciencedirect-ai, 10 February 2026) were utilized exclusively for language refinement, grammar correction, academic tone enhancement, and manuscript formatting. DeepL Write provided sentence-level writing improvements and clarity enhancements. ScienceDirect AI assisted with literature search refinement and related terminology accuracy. All research design, methodology,

data analysis, result interpretation, and scientific conclusions were entirely conceived, executed, and verified by the author. Following application of these tools, the manuscript was thoroughly reviewed and edited to ensure all content reflects the author's original work and scientific intent. The author assumes full responsibility for the accuracy, integrity, and scientific validity of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ANN	Artificial Neural Network
CCS	Carbon Capture and Storage
CI	Confidence Interval
CO_2	Carbon Dioxide
CPU	Central Processing Unit
CSVs	Comma Separated Values
DT	Digital Twin
GBM	Gradient Boosting Machine
GPU	Graphics Processing Unit
L/G	Liquid-to-Gas Ratio
LHS	Latin Hypercube Sampling
MAE	Mean Absolute Error
MAPD	Mean Absolute Percentage Deviation
MEA	Monoethanolamine
ML	Machine Learning
MPC	Model Predictive Control
NSGA-II	Non-dominated Sorting Genetic Algorithm II
NRTL	Non-Random Two Liquid (activity coefficient model)
RBF	Radial Basis Function
RMSE	Root Mean Square Error
RMSPD	Root Mean Square Percentage Deviation
R^2	Coefficient of Determination
SRD	Specific Regeneration Duty
SVR	Support Vector Regression
VLE	Vapor–Liquid Equilibrium
XGBoost	Extreme Gradient Boosting
SHAP	SHapley Additive exPlanations

References

1. Intergovernmental Panel on Climate Change (IPCC). *Climate Change 2021—The Physical Science Basis*; Cambridge University Press: Cambridge, UK, 2023; ISBN 9781009157896.
2. Rabbi, M.F.; Abdullah, M. Fossil Fuel CO_2 Emissions and Economic Growth in the Visegrád Region: A Study Based on the Environmental Kuznets Curve Hypothesis. *Climate* **2024**, *12*, 115. [CrossRef]
3. Pianta, S.; Rinscheid, A.; Weber, E.U. Carbon Capture and Storage in the United States: Perceptions, Preferences, and Lessons for Policy. *Energy Policy* **2021**, *151*, 112149. [CrossRef]
4. Rabbi, M.F. Unified Artificial Intelligence Framework for Modeling Pollution Dynamics and Sustainable Remediation in Environmental Chemistry. *Sci. Rep.* **2025**, *15*, 36196. [CrossRef]
5. Wang, Y.; Pan, Z.; Zhang, W.; Borhani, T.N.; Li, R.; Zhang, Z. Life Cycle Assessment of Combustion-Based Electricity Generation Technologies Integrated with Carbon Capture and Storage: A Review. *Environ. Res.* **2022**, *207*, 112219. [CrossRef]
6. van Ruijven, B.J.; van Vuuren, D.P.; Boskaljon, W.; Neelis, M.L.; Saygin, D.; Patel, M.K. Long-Term Model-Based Projections of Energy Use and CO_2 Emissions from the Global Steel and Cement Industries. *Resour. Conserv. Recycl.* **2016**, *112*, 15–36. [CrossRef]
7. Zhang, Z.; Vo, D.-N.; Nguyen, T.B.H.; Sun, J.; Lee, C.-H. Advanced Process Integration and Machine Learning-Based Optimization to Enhance Techno-Economic-Environmental Performance of CO_2 Capture and Conversion to Methanol. *Energy* **2024**, *293*, 130758. [CrossRef]

8. Wennersten, R.; Sun, Q.; Li, H. The Future Potential for Carbon Capture and Storage in Climate Change Mitigation—An Overview from Perspectives of Technology, Economy and Risk. *J. Clean. Prod.* **2015**, *103*, 724–736. [CrossRef]
9. Gidden, M.J.; Joshi, S.; Armitage, J.J.; Christ, A.-B.; Boettcher, M.; Brutschin, E.; Köberle, A.C.; Riahi, K.; Schellnhuber, H.J.; Schleussner, C.-F.; et al. A Prudent Planetary Limit for Geologic Carbon Storage. *Nature* **2025**, *645*, 124–132. [CrossRef]
10. Rubin, E.S.; Davison, J.E.; Herzog, H.J. The Cost of CO_2 Capture and Storage. *Int. J. Greenh. Gas Control* **2015**, *40*, 378–400. [CrossRef]
11. Gingerich, D.B.; Grol, E.; Mauter, M.S. Fundamental Challenges and Engineering Opportunities in Flue Gas Desulfurization Wastewater Treatment at Coal Fired Power Plants. *Environ. Sci.* **2018**, *4*, 909–925. [CrossRef]
12. Wang, H.; Liu, S.; Wang, H.; Chao, J.; Li, T.; Ellis, N.; Duo, W.; Bi, X.; Smith, K.J. Thermochemical Conversion of Biomass to Fuels and Chemicals: A Review of Catalysts, Catalyst Stability, and Reaction Mechanisms. *Catal. Rev.* **2025**, *67*, 57–129. [CrossRef]
13. Tahir, N.M.; Zhang, J.; Armstrong, M. Control of Heat-Integrated Distillation Columns: Review, Trends, and Challenges for Future Research. *Processes* **2024**, *13*, 17. [CrossRef]
14. Mommers, J.; van der Wal, S. Column Selection and Optimization for Comprehensive Two-Dimensional Gas Chromatography: A Review. *Crit. Rev. Anal. Chem.* **2021**, *51*, 183–202. [CrossRef] [PubMed]
15. Zhang, Q.; Turton, R.; Bhattacharyya, D. Development of Model and Model-Predictive Control of an MEA-Based Postcombustion CO_2 Capture Process. *Ind. Eng. Chem. Res.* **2016**, *55*, 1292–1308. [CrossRef]
16. Li, Q.; Zhang, W.; Qin, Y.; An, A. Model Predictive Control for the Process of MEA Absorption of CO_2 Based on the Data Identification Model. *Processes* **2021**, *9*, 183. [CrossRef]
17. Harun, N.; Nittaya, T.; Douglas, P.L.; Croiset, E.; Ricardez-Sandoval, L.A. Dynamic Simulation of MEA Absorption Process for CO_2 Capture from Power Plants. *Int. J. Greenh. Gas Control* **2012**, *10*, 295–309. [CrossRef]
18. Ghadyanlou, F.; Azari, A.; Vatani, A. A Review of Modeling Rotating Packed Beds and Improving Their Parameters: Gas–Liquid Contact. *Sustainability* **2021**, *13*, 8046. [CrossRef]
19. Tobiesen, F.A.; Svendsen, H.F.; Juliussen, O. Experimental Validation of a Rigorous Absorber Model for CO_2 Postcombustion Capture. *AIChE J.* **2007**, *53*, 846–865. [CrossRef]
20. Wu, X.; Wang, M.; Shen, J.; Li, Y.; Lawal, A.; Lee, K.Y. Reinforced Coordinated Control of Coal-Fired Power Plant Retrofitted with Solvent Based CO_2 Capture Using Model Predictive Controls. *Appl. Energy* **2019**, *238*, 495–515. [CrossRef]
21. Bartling, A.W.; Stone, M.L.; Hanes, R.J.; Bhatt, A.; Zhang, Y.; Biddy, M.J.; Davis, R.; Kruger, J.S.; Thornburg, N.E.; Luterbacher, J.S.; et al. Techno-Economic Analysis and Life Cycle Assessment of a Biorefinery Utilizing Reductive Catalytic Fractionation. *Energy Environ. Sci.* **2021**, *14*, 4147–4168. [CrossRef]
22. Sipöcz, N.; Tobiesen, F.A.; Assadi, M. The Use of Artificial Neural Network Models for CO_2 Capture Plants. *Appl. Energy* **2011**, *88*, 2368–2376. [CrossRef]
23. Fu, J.; Chang, Y.; Huang, B. Prediction and Sensitivity Analysis of CO_2 Capture by Amine Solvent Scrubbing Technique Based on BP Neural Network. *Front. Bioeng. Biotechnol.* **2022**, *10*, 907904. [CrossRef] [PubMed]
24. Hu, X.; Guo, L.; Wang, J.; Liu, Y. Computational Fluid Dynamics and Machine Learning Integration for Evaluating Solar Thermal Collector Efficiency -Based Parameter Analysis. *Sci. Rep.* **2025**, *15*, 24528. [CrossRef] [PubMed]
25. Yan, Y.; Borhani, T.N.; Subraveti, S.G.; Pai, K.N.; Prasad, V.; Rajendran, A.; Nkulikiyinka, P.; Asibor, J.O.; Zhang, Z.; Shao, D.; et al. Harnessing the Power of Machine Learning for Carbon Capture, Utilisation, and Storage (CCUS)—A State-of-the-Art Review. *Energy Environ. Sci.* **2021**, *14*, 6122–6157. [CrossRef]
26. Manikandan, S.; Kaviya, R.S.; Shreeharan, D.H.; Subbaiya, R.; Vickram, S.; Karmegam, N.; Kim, W.; Govarthanan, M. Artificial Intelligence-driven Sustainability: Enhancing Carbon Capture for Sustainable Development Goals—A Review. *Sustain. Dev.* **2025**, *33*, 2004–2029. [CrossRef]
27. Ebadian, M.; van Dyk, S.; McMillan, J.D.; Saddler, J. Biofuels Policies That Have Encouraged Their Production and Use: An International Perspective. *Energy Policy* **2020**, *147*, 111906. [CrossRef]
28. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
29. Venkatasubramanian, V. The Promise of Artificial Intelligence in Chemical Engineering: Is It Here, Finally? *AIChE J.* **2019**, *65*, 466–478. [CrossRef]
30. Sun, C.; Liu, Z.-P. Discovering Explainable Biomarkers for Breast Cancer Anti-PD1 Response via Network Shapley Value Analysis. *Comput. Methods Programs Biomed.* **2024**, *257*, 108481. [CrossRef]
31. Tan, S.; Wang, R.; Song, G.; Qi, S.; Zhang, K.; Zhao, Z.; Yin, Q. Machine Learning and Shapley Additive Explanation-Based Interpretable Prediction of the Electrocatalytic Performance of N-Doped Carbon Materials. *Fuel* **2024**, *355*, 129469. [CrossRef]
32. Kumar De, S.; Won, D.-I.; Kim, J.; Kim, D.H. Integrated CO_2 Capture and Electrochemical Upgradation: The Underpinning Mechanism and Techno-Chemical Analysis. *Chem. Soc. Rev.* **2023**, *52*, 5744–5802. [CrossRef]

33. Nguyen, A.-T.; Reiter, S.; Rigo, P. A Review on Simulation-Based Optimization Methods Applied to Building Performance Analysis. *Appl. Energy* **2014**, *113*, 1043–1058. [CrossRef]
34. Alyahya, A.; Lannon, S.; Jabi, W. A Framework for Optimizing Biomimetic Opaque Ventilated Façades Using CFD and Machine Learning. *Buildings* **2025**, *15*, 4130. [CrossRef]
35. Bhosekar, A.; Ierapetritou, M. Advances in Surrogate Based Modeling, Feasibility Analysis, and Optimization: A Review. *Comput. Chem. Eng.* **2018**, *108*, 250–267. [CrossRef]
36. Nian, M.; Dong, R.; Zhong, W.; Zhang, Y.; Lou, D. Multi-Objective Optimization Study on Capture Performance of Diesel Particulate Filter Based on the GRA-MLR-WOA Hybrid Method. *Sustainability* **2025**, *17*, 8777. [CrossRef]
37. Yang, Y.; Li, Y.; Huang, Q.; Xia, J.; Li, J. Surrogate-Based Multiobjective Optimization to Rapidly Size Low Impact Development Practices for Outflow Capture. *J. Hydrol.* **2023**, *616*, 128848. [CrossRef]
38. Omrany, H.; Al-Obaidi, K.M.; Husain, A.; Ghaffarianhoseini, A. Digital Twins in the Construction Industry: A Comprehensive Review of Current Implementations, Enabling Technologies, and Future Directions. *Sustainability* **2023**, *15*, 10908. [CrossRef]
39. Rasheed, A.; San, O.; Kvamsdal, T. Digital Twin: Values, Challenges and Enablers From a Modeling Perspective. *IEEE Access* **2020**, *8*, 21980–22012. [CrossRef]
40. Rahimi, M.; Moosavi, S.M.; Smit, B.; Hatton, T.A. Toward Smart Carbon Capture with Machine Learning. *Cell Rep. Phys. Sci.* **2021**, *2*, 100396. [CrossRef]
41. Chung, W.; Lee, J.H. Input–Output Surrogate Models for Efficient Economic Evaluation of Amine Scrubbing CO_2 Capture Processes. *Ind. Eng. Chem. Res.* **2020**, *59*, 18951–18964. [CrossRef]
42. von Stosch, M.; Oliveira, R.; Peres, J.; Feyo de Azevedo, S. Hybrid Semi-Parametric Modeling in Process Systems Engineering: Past, Present and Future. *Comput. Chem. Eng.* **2014**, *60*, 86–101. [CrossRef]
43. Motaei, E.; Ganat, T.; Tabatabai, M.; Umer, M.; Krishna, S. Advancements in Machine Learning–Driven Proxy Modelling Techniques for Reservoir Engineering: A Systematic Review. *J. Pet. Geol.* **2026**, *49*, 516–540. [CrossRef]
44. Gao, X.; Chen, B.; He, X.; Qiu, T.; Li, J.; Wang, C.; Zhang, L. Multi-Objective Optimization for the Periodic Operation of the Naphtha Pyrolysis Process Using a New Parallel Hybrid Algorithm Combining NSGA-II with SQP. *Comput. Chem. Eng.* **2008**, *32*, 2801–2811. [CrossRef]
45. Ziaii, S.; Rochelle, G.T.; Edgar, T.F. Optimum Design and Control of Amine Scrubbing in Response to Electricity and CO_2 Prices. *Energy Procedia* **2011**, *4*, 1683–1690. [CrossRef]
46. Bailey, N.; Papakyriakou, T.N.; Bartels, C.; Wang, F. Henry's Law Constant for CO_2 in Aqueous Sodium Chloride Solutions at 1 Atm and Sub-Zero (Celsius) Temperatures. *Mar. Chem.* **2018**, *207*, 26–32. [CrossRef]
47. Gladis, A.; Lomholdt, N.F.; Fosbøl, P.L.; Woodley, J.M.; von Solms, N. Pilot Scale Absorption Experiments with Carbonic Anhydrase-Enhanced MDEA- Benchmarking with 30 wt% MEA. *Int. J. Greenh. Gas Control* **2019**, *82*, 69–85. [CrossRef]
48. Hu, Y.; Wang, Q.; Hu, D.; Zhang, Y.; Furqan, M.; Lu, S. Experimental Study on CO_2 Capture by MEA/n-Butanol/H_2O Phase Change Absorbent. *RSC Adv.* **2024**, *14*, 3146–3157. [CrossRef]
49. Vinjarapu, S.H.B.; Neerup, R.; Larsen, A.H.; Jørsboe, J.K.; Villadsen, S.N.B.; Jensen, S.; Karlsson, J.L.; Kappel, J.; Lassen, H.; Blinksbjerg, P.; et al. Results from Pilot-Scale CO Capture Testing Using 30 wt% MEA at a Waste-to-Energy Facility: Optimisation through Parametric Analysis. *Appl. Energy* **2024**, *355*, 122193. [CrossRef]
50. McKay, M.D.; Beckman, R.J.; Conover, W.J. Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **1979**, *21*, 239–245. [CrossRef]
51. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874. [CrossRef]
52. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef]
53. Mangalapally, H.P.; Notz, R.; Hoch, S.; Asprion, N.; Sieder, G.; Garcia, H.; Hasse, H. Pilot Plant Experimental Studies of Post Combustion CO_2 Capture by Reactive Absorption with MEA and New Solvents. *Energy Procedia* **2009**, *1*, 963–970. [CrossRef]
54. Bui, M.; Adjiman, C.S.; Bardow, A.; Anthony, E.J.; Boston, A.; Brown, S.; Fennell, P.S.; Fuss, S.; Galindo, A.; Hackett, L.A.; et al. Carbon Capture and Storage (CCS): The Way Forward. *Energy Environ. Sci.* **2018**, *11*, 1062–1176. [CrossRef]
55. Nittaya, T.; Douglas, P.L.; Croiset, E.; Ricardez-Sandoval, L.A. Dynamic Modelling and Control of MEA Absorption Processes for CO_2 Capture from Power Plants. *Fuel* **2014**, *116*, 672–691. [CrossRef]
56. Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Machine Learning Models Using Shapley Values: Application to Compound Potency and Multi-Target Activity Predictions. *J. Comput. Aided. Mol. Des.* **2020**, *34*, 1013–1026. [CrossRef]
57. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [CrossRef]
58. Helton, J.C.; Davis, F.J. Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems. *Reliab. Eng. Syst. Saf.* **2003**, *81*, 23–69. [CrossRef]

59. Lawal, A.; Wang, M.; Stephenson, P.; Obi, O. Demonstrating Full-Scale Post-Combustion CO_2 Capture for Coal-Fired Power Plants through Dynamic Modelling and Simulation. *Fuel* **2012**, *101*, 115–128. [CrossRef]
60. Bove, C.; Laugel, T.; Lesot, M.-J.; Tijus, C.; Detyniecki, M. Why Do Explanations Fail? A Typology and Discussion on Failures in XAI. *arXiv* **2025**, arXiv:2405.13474.