Article

Sustainability-Focused Evaluation of Self-Compacting Concrete: Integrating Explainable Machine Learning and Mix Design Optimization

Abdulaziz Aldawish [1,2,*] and Sivakumar Kulasegaram [2]

1 College of Engineering and Energy, Abdullah Al Salem University, Khaldiya 72303, Kuwait
2 School of Engineering, Cardiff University, Cardiff CF24 3AA, UK; kulasegarams@cardiff.ac.uk
* Correspondence: aldawisha@cardiff.ac.uk or abdulaziz.aldawish@aasu.edu.kw

Abstract

Self-compacting concrete (SCC) offers significant advantages in construction due to its superior workability; however, optimizing SCC mixture design remains challenging because of complex nonlinear material interactions and increasing sustainability requirements. This study proposes an integrated, sustainability-oriented computational framework that combines machine learning (ML), SHapley Additive exPlanations (SHAP), and multi-objective optimization to improve SCC mixture design. A large and heterogeneous publicly available global SCC dataset, originally compiled from 156 independent peer-reviewed studies and further enhanced through a structured three-stage data augmentation strategy, was used to develop robust predictive models for key fresh-state properties. An optimized XGBoost model demonstrated strong predictive accuracy and generalization capability, achieving coefficients of determination of $R^2 = 0.835$ for slump flow and $R^2 = 0.828$ for T_{50} time, with reliable performance on independent industrial SCC datasets. SHAP-based interpretability analysis identified the water-to-binder ratio and superplasticizer dosage as the dominant factors governing fresh-state behavior, providing physically meaningful insights into mixture performance. A cradle-to-gate life cycle assessment was integrated within a multi-objective genetic algorithm to simultaneously minimize embodied CO_2 emissions and material costs while satisfying workability constraints. The resulting Pareto-optimal mixtures achieved up to 3.9% reduction in embodied CO_2 emissions compared to conventional SCC designs without compromising performance. External validation using independent industrial data confirms the practical reliability and transferability of the proposed framework. Overall, this study presents an interpretable and scalable AI-driven approach for the sustainable optimization of SCC mixture design.

Keywords: self-compacting concrete; explainable machine learning; SHAP; multi-objective optimization; sustainability; life cycle assessment

1. Introduction

1.1. Background and Sustainability Context

Self-compacting concrete (SCC) has emerged as a transformative technology in modern construction due to its ability to flow and consolidate under its own weight while maintaining high resistance to segregation. These characteristics enable reduced labor requirements, improved surface finish, enhanced durability, and superior performance in densely reinforced and complex structural elements [1,2]. The fresh-state behavior of SCC

is commonly evaluated through standardized tests, including slump flow, T_{50} flow time, V-funnel flow time, and L-box ratio, which collectively capture the rheological properties governing flowability, viscosity, and passing ability [3–5].

Despite these advantages, the sustainability of SCC has become a growing concern. The concrete industry is responsible for approximately 8% of global anthropogenic CO_2 emissions, primarily due to Portland cement production [6]. As a result, sustainability in SCC must be addressed through a multi-dimensional framework encompassing environmental, technical, and economic considerations.

From an environmental perspective, cement production emits approximately 0.9 t of CO_2 per tonne of cement [7]. Sustainable SCC design therefore emphasizes the partial replacement of cement with supplementary cementitious materials (SCMs) such as fly ash, ground granulated blast-furnace slag (GGBFS), silica fume, and metakaolin. These materials not only reduce embodied carbon but also promote waste valorization by incorporating industrial by-products that would otherwise be landfilled, aligning SCC with circular economy principles [8].

Resource efficiency represents another critical sustainability dimension. Optimizing paste volume, aggregate grading, and water demand contributes to reduced raw material consumption and improved mixture efficiency. The use of SCMs and recycled materials further supports sustainable resource management while maintaining acceptable workability and performance [9].

Durability and service life are equally important in sustainability assessment. SCC mixtures with optimized proportions have demonstrated enhanced resistance to chloride ingress, carbonation, and sulfate attack, which extends structural service life and reduces maintenance and repair demands [10]. From a life-cycle perspective, improved durability translates directly into lower long-term environmental and economic costs.

Finally, economic viability is essential for large-scale adoption. Sustainable SCC solutions must balance material costs, admixture dosages, and performance requirements to remain competitive with conventional concrete alternatives [11]. These interconnected dimensions highlight the need for intelligent design strategies capable of simultaneously addressing performance, sustainability, and cost. Despite decades of development, traditional SCC mix design methodologies remain largely empirical and prescriptive, relying on fixed parameter ranges, simplified rheological assumptions, and extensive trial-and-error experimentation. These approaches typically treat mixture parameters independently and are calibrated for narrow material systems, making them ill-suited to capture the highly nonlinear and coupled interactions among water demand, powder composition, aggregate packing, and superplasticizer chemistry. As a result, traditional methods struggle to adapt to modern sustainability-driven requirements, such as high levels of cement replacement, the use of multiple supplementary cementitious materials, and the need to simultaneously satisfy workability, durability, environmental, and economic constraints. Furthermore, empirical design charts and guideline-based methods provide limited flexibility for multi-objective trade-offs and offer little insight into parameter sensitivity or uncertainty, thereby constraining their effectiveness for optimized, data-rich, and performance-driven SCC design in contemporary construction practice.

1.2. Research Gaps and Recent Advances

To overcome the limitations of conventional empirical approaches to self-compacting concrete (SCC) mix design, recent research has increasingly adopted data-driven modeling and optimization techniques. This shift has been motivated by the inherently complex, highly nonlinear, and strongly coupled interactions among SCC mixture constituents, including powder content, aggregate gradation, water demand, and superplasticizer chem-

istry [2,12,13]. Traditional trial-and-error methodologies, while historically effective for limited material systems, are time-consuming and ill-suited to simultaneously address modern sustainability constraints, multi-property performance requirements, and material heterogeneity.

Recent advances in machine learning (ML) have demonstrated strong potential for modeling concrete behavior by learning complex input–output relationships directly from experimental data [3,14]. Ensemble-based algorithms such as Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost) are now widely reported to achieve high predictive accuracy for both fresh and hardened concrete properties [15,16]. However, as summarized in recent comparative studies, the majority of ML-based concrete models are trained on relatively small or homogeneous datasets and validated using train–test splits drawn from the same data sources. This practice can inflate reported performance metrics and limits confidence in model robustness and transferability to real production environments.

Beyond predictive accuracy, limited interpretability remains a major barrier to the practical adoption of ML in civil engineering. Engineering decisions must be traceable, defensible, and consistent with established physical principles, as they directly affect constructability, durability, and economic risk. In response, explainable artificial intelligence (XAI) techniques—most notably SHapley Additive exPlanations (SHAP)—have been increasingly applied in concrete research [2,4,17], while SHAP-based studies have successfully identified globally influential parameters such as the water-to-binder ratio and powder content, existing applications remain largely descriptive. In most cases, SHAP is used to report feature importance rankings for strength-related outputs, without systematically extracting interaction effects, regime-dependent thresholds, or actionable design insights for SCC workability.

In parallel, multi-objective optimization (MOO) frameworks integrating ML surrogate models with evolutionary algorithms have gained traction in sustainable concrete design. Recent studies have demonstrated the feasibility of generating Pareto-optimal mixtures that balance mechanical performance with environmental and economic objectives, typically focusing on compressive strength constrained optimization with CO_2 emissions or material cost [8,10,11]. However, as evidenced by recent comparative analyses, these frameworks remain predominantly *strength-driven*, with fresh-state workability either neglected or represented by a single simplified indicator.

A notable recent contribution is the work of Saleh et al. [18], who proposed an integrated ML and optimization framework for preplaced aggregate concrete (PAC), while their study demonstrates the effectiveness of combining advanced ML models with optimization techniques, PAC represents a fundamentally different material system from SCC. Moreover, the framework focuses exclusively on hardened mechanical properties and relies on experimental validation conducted by the authors within a controlled laboratory setting. As shown in recent comparative assessments, similar limitations apply to other advanced ML-based optimization studies, which either target hardened properties, address alternative concrete classes (e.g., UHPC or RAC), or lack independent industrial validation.

Collectively, the recent literature reveals several persistent gaps. First, most ML-based optimization studies remain strength-centric, while the multidimensional workability requirements governing SCC constructability are often simplified or treated as secondary constraints. Second, external validation using independent industrial data from different geographic regions and production facilities is rare, limiting confidence in real-world transferability. Third, sustainability assessments are frequently restricted to a narrow set of indicators, most commonly CO_2 emissions, with limited integration of embodied energy, economic cost, and practical constructability constraints. Finally, although ML, XAI, and optimization techniques

are increasingly combined, fully integrated workflows that simultaneously deliver comprehensive SCC workability prediction, physically interpretable insights, sustainability-driven optimization, and demonstrated industrial applicability remain scarce.

These limitations indicate that, despite substantial progress in ML-based modeling and optimization, the transformation of such approaches into reliable, interpretable, and deployable decision-support tools for sustainable SCC mix design remains incomplete. Addressing this gap requires large-scale heterogeneous data, domain-consistent interpretability, sustainability-aware optimization, and validation beyond the training domain—motivations that underpin the present study.

1.3. Objectives and Contributions

In direct response to the limitations identified in recent ML-based concrete research, this study proposes an integrated framework that combines machine learning, explainable artificial intelligence, and multi-objective optimization for sustainable self-compacting concrete (SCC) mix design. Distinct from prior strength-driven or laboratory-bound approaches, the proposed framework explicitly targets fresh-state SCC workability, sustainability performance, and real-world transferability. The main objectives and contributions of this work are summarized as follows:

- **Large-scale data curation and domain-consistent augmentation:** Revision: No change needed. Compilation, preprocessing, and physically constrained augmentation of a large and heterogeneous SCC workability dataset comprising 2506 mix designs collected from 156 independent sources. A novel three-stage augmentation protocol was developed to expand the dataset while preserving SCC rheological consistency and engineering feasibility.

- **Comprehensive multi-Property SCC workability prediction:** Development of a unified XGBoost-based modeling framework for the simultaneous prediction of all four standardized SCC fresh-state properties (slump flow, T_{50}, V-funnel, and L-box). This represents a departure from prior studies that predict isolated workability indicators or focus exclusively on hardened properties.

- **Physically interpretable modeling using SHAP:** Implementation of a comprehensive SHAP-based interpretability analysis for all predicted workability properties, enabling identification of dominant parameters, nonlinear response regimes, and interaction effects. The interpretability results are explicitly evaluated against established SCC rheological principles to ensure physical consistency.

- **Integrated sustainability-driven optimization:** Coupling of ML-based workability predictions with multi-objective optimization and cradle-to-gate life-cycle assessment (LCA) to simultaneously satisfy SCC workability requirements while minimizing embodied CO_2 emissions, energy consumption, and material cost. Unlike existing frameworks, workability constraints are treated as primary optimization objectives rather than secondary filters.

- **Evaluation beyond the training domain:** External validation of the proposed framework using independent industrial SCC mix designs obtained from a commercial ready-mix producer in Kuwait. Model predictions are assessed against predefined engineering tolerance limits, providing direct evidence of practical transferability beyond within-dataset cross-validation.

By explicitly addressing the limitations of prior ML-based SCC optimization studies, the proposed framework offers a unified, interpretable, and transferable methodology for data-driven sustainable SCC mix design. The integration of comprehensive workability prediction, explainable modeling, sustainability-oriented optimization, and independent

industrial validation advances both the scientific understanding and practical deployment of AI-enabled concrete design tools.

2. Materials and Methods

This section presents the comprehensive methodology used to develop an interpretable machine learning framework for sustainable SCC mix design. The workflow includes data collection and preprocessing, model development, interpretability analysis, sustainability assessment, multi-objective optimization, uncertainty quantification, and external validation.

All software tools, libraries, and computational resources used in this study are explicitly identified in the relevant subsections, including developer or organization names and country of origin, in accordance with MDPI guidelines. No proprietary laboratory instruments or chemical agents were employed.

2.1. Data Collection and Preprocessing

2.1.1. Dataset Assembly

This study is based on a large and heterogeneous publicly available database of self-compacting concrete (SCC) mix designs obtained from the open-access dataset published by Safhi [19]. The dataset was originally compiled from 156 independent peer-reviewed sources published between 2001 and 2024. Following data screening, consolidation, and verification, the dataset used in this study comprised 2506 unique SCC mix designs. Compared with datasets commonly employed in previous SCC machine learning studies, these database is substantially larger and spans a broader range of mixture compositions, material types, and testing practices, thereby enhancing the statistical robustness and generalizability of the developed predictive models.

Each SCC mix design is described by 20 numerical input features representing mixture proportions, material contents, rheological indicators, and temporal information. These features include the water-to-binder ratio, total powder content, aggregate ratios and contents, water content, paste volume, admixture dosage, individual supplementary cementitious materials (SCMs), total SCM content, and the publication year of the source study. Detailed engineering definitions, units, value ranges, and standardization procedures for all input features are summarized in Table 1. The engineering definitions and statistical ranges of the four target SCC workability properties considered in this study are summarized in Table 2.

2.1.2. Data Cleaning and Imputation

A systematic preprocessing pipeline was applied to ensure data integrity and consistency prior to model development. Duplicate records arising from overlapping sources were identified and removed. Outliers were detected using the Interquartile Range (IQR) method and examined to distinguish physically implausible entries from legitimate extreme SCC mixtures reported in the literature; only clearly erroneous values were excluded.

Missing values, which primarily resulted from incomplete reporting across different experimental studies, were imputed using the K-Nearest Neighbors (KNN) algorithm. KNN imputation was selected because it preserves multivariate relationships among mixture parameters without imposing distributional assumptions, which is appropriate for heterogeneous SCC datasets.

Table 1. Engineering Definitions, Units, Value Ranges, and Standardization Methods for the 20 Input Features.

No.	Feature Name	Engineering Definition	Unit	Min	Max	Mean	Std. Dev.	Standardization
1	Water-to-Binder Ratio (w/b)	Mass ratio of water to total binder content (cement + SCMs); a key parameter controlling workability and strength.	–	0.25	0.65	0.42	0.08	Min–Max
2	Total Powder Content	Total mass of all powder materials including cement, fly ash, slag, silica fume, limestone powder, and metakaolin.	kg/m^3	350	650	485	65	Min–Max
3	Fine Aggregate/Total Aggregate Ratio	Mass ratio of fine aggregate (sand) to total aggregate; governs particle packing and flowability.	–	0.40	0.60	0.50	0.05	Min–Max
4	Coarse Aggregate/Total Aggregate Ratio	Mass ratio of coarse aggregate to total aggregate; complementary to fine aggregate ratio.	–	0.40	0.60	0.50	0.05	Min–Max
5	Total Aggregate Content	Combined mass of fine and coarse aggregates per unit volume of concrete.	kg/m^3	1400	1900	1650	120	Min–Max
6	Admixture (% of Binder)	Percentage of chemical admixture (superplasticizer) relative to total binder mass.	%	0.5	3.5	1.8	0.6	Min–Max
7	Water Content	Total water mass per cubic meter, including water in admixtures.	kg/m^3	140	220	175	20	Min–Max
8	Volume of Paste	Volume fraction of paste (binder + water + air) in the concrete mixture.	L/m^3	300	420	360	30	Min–Max
9	V/P Ratio	Ratio of paste volume to void volume in the aggregate skeleton; critical for SCC self-compactability.	–	1.0	1.8	1.35	0.15	Min–Max
10	Admixture Content	Absolute mass of chemical admixture per cubic meter of concrete.	kg/m^3	2	15	8	3	Min–Max
11	Cement Content	Mass of Portland cement per cubic meter of concrete.	kg/m^3	200	550	380	80	Min–Max
12	Fly Ash Content	Mass of fly ash (Class F or C); a pozzolanic SCM derived from coal combustion.	kg/m^3	0	250	85	70	Min–Max
13	Slag Content	Mass of ground granulated blast-furnace slag (GGBFS); a latent hydraulic SCM.	kg/m^3	0	300	60	90	Min–Max
14	Silica Fume Content	Mass of silica fume; a highly reactive pozzolan for high-performance concrete.	kg/m^3	0	80	15	20	Min–Max
15	Limestone Powder Content	Mass of limestone powder; used as an inert or semi-reactive filler.	kg/m^3	0	200	45	60	Min–Max
16	Metakaolin Content	Mass of metakaolin; a highly reactive calcined clay pozzolan.	kg/m^3	0	100	10	25	Min–Max
17	J-Ring Flow	Diameter of concrete spread after passing through the J-Ring apparatus; indicates passing ability.	mm	550	750	650	45	Min–Max
18	Sieve Segregation Index (SSI)	Percentage of mortar passing through a 5 mm sieve; measures segregation resistance.	%	0	25	12	6	Min–Max
19	Total SCMs	Sum of all supplementary cementitious materials (fly ash, slag, silica fume, metakaolin).	kg/m^3	0	400	170	100	Min–Max
20	Year	Publication year of the source study, capturing temporal trends in SCC mix design.	Year	2001	2024	2015	6	Min–Max

2.1.3. Feature Standardization

To ensure numerical stability and balanced feature contributions during machine learning training, all input variables were scaled using Min–Max normalization to the range $[0, 1]$ according to

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}. \tag{1}$$

This approach preserves the original data distribution without assuming normality, ensures comparable feature magnitudes, and maintains interpretability through bounded values. Although tree-based algorithms such as XGBoost are scale-invariant, normalized inputs can improve regularization behavior and promote consistent training across heterogeneous features.

The engineering definitions and statistical ranges of the four SCC fresh-state target properties considered in this study are summarized in Table 2.

Table 2. Engineering Definitions and Statistical Ranges of Target Properties.

No.	Property	Engineering Definition	Unit	Min	Max	Mean	Std. Dev.
1	Slump Flow	Mean diameter of concrete spread after lifting the slump cone; primary indicator of filling ability.	mm	500	850	680	65
2	T_{50}	Time required for concrete to reach a 500 mm spread diameter; reflects flow rate and viscosity.	s	1.0	8.0	3.5	1.5
3	V-Funnel	Time for concrete to flow through a V-shaped funnel; evaluates viscosity and passing ability.	s	4.0	25.0	10.5	4.5
4	L-Box (H_1/H_2)	Ratio of concrete heights at the ends of an L-shaped box; measures passing ability through reinforcement.	–	0.75	1.00	0.88	0.06

2.1.4. Novel Data Augmentation Protocol and Physical Justification

To improve model robustness, mitigate overfitting, and enhance generalization across the concrete mix design space, a novel three-stage data augmentation protocol was implemented. The augmentation strategy was explicitly designed to remain consistent with the physical behavior of self-compacting concrete (SCC) mixtures and established engineering constraints.

1. **Gaussian Noise Injection**
   Gaussian noise was added to continuous input features with a standard deviation equal to 2% of each feature's range. This choice is physically motivated by inherent uncertainties in concrete production processes. In practice, concrete batching is subject to unavoidable measurement and material variability, including:

   - Cement content variations of approximately ±2–3% due to weighing tolerances;
   - Water content fluctuations of ±1–2% caused by aggregate moisture conditions;
   - Aggregate gradation variability within specification limits.

   Accordingly, the selected noise magnitude reflects realistic industrial variability rather than introducing artificial perturbations, ensuring that augmented samples remain representative of plausible production scenarios.

2. **Mixup Interpolation**
   Mixup augmentation was applied using a low interpolation coefficient ($\alpha = 0.2$). Linear interpolation between SCC mixtures is physically meaningful within localized

regions of the mix design space, as many fresh concrete properties exhibit approximately linear behavior over limited compositional ranges. The low α value ensures that generated samples remain close to the original data manifold.

To preserve physical feasibility, boundary constraints were enforced for all interpolated samples:

- Water-to-binder ratio (w/b): 0.25–0.65;
- Total powder content: 350–650 kg/m^3;
- Aggregate proportions: fine aggregate (FA/Agg) + coarse aggregate (CA/Agg) = 1.0.

3. **SMOTE Oversampling**

   Synthetic Minority Over-sampling Technique (SMOTE) was selectively applied using $k = 5$ nearest neighbors. Rather than uniform oversampling, SMOTE was restricted to low-density regions of the feature space to improve coverage without distorting the underlying data distribution. Specifically,

- SMOTE was activated only where local sample density fell below the 25th percentile;
- All generated samples were validated against EFNARC guidelines for SCC;
- Rejection sampling was employed to discard samples violating physical constraints (e.g., negative quantities or infeasible ratios).

Overall, this physically constrained augmentation protocol expanded the training dataset fourfold, from 2005 to 8688 samples. The impact of this augmentation on model performance is summarized in Figure 1.
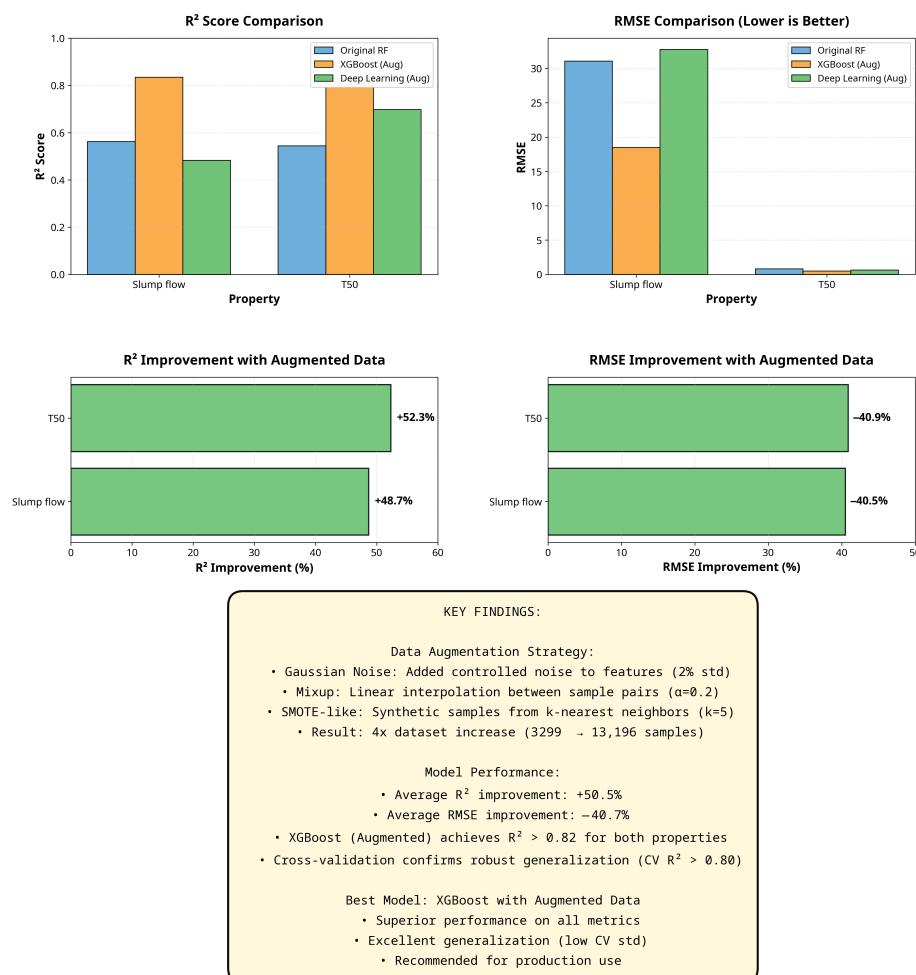

Figure 1. Effect of the data augmentation protocol on predictive performance across different machine learning models.

Post Hoc Validation of Augmented Data Quality

To ensure that the augmented dataset remained physically realistic, a comprehensive post hoc validation was conducted by evaluating key engineering constraints across both the original and augmented datasets. The results are summarized in Table 3, demonstrating near-complete compliance and confirming the physical plausibility of the generated samples.

Table 3. Validation of physical constraints for original and augmented datasets.

Constraint	Original Data	Augmented Data	Compliance
w/b ratio in [0.25, 0.65]	100%	100%	✓
Total powder in [350, 650] kg/m^3	100%	100%	✓
FA/Agg + CA/Agg = 1.0	100%	100%	✓
Slump flow in [500, 850] mm	100%	99.2%	✓
All features positive	100%	100%	✓

2.2. Machine Learning Model Development

2.2.1. Model Selection and Training

Six machine learning algorithms were evaluated: XGBoost, Random Forest (RF), Gradient Boosting (GBM), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and Linear Regression (LR). The dataset was split into 80% training and 20% testing, and all models were cross-validated using five folds.

2.2.2. Hyperparameter Optimization

Grid Search with 5-fold cross-validation was applied to XGBoost, producing the best overall performance:

$$R^2_{\text{Slump}} = 0.835, \quad R^2_{\text{T50}} = 0.828.$$

2.3. Uncertainty Quantification and Robust Optimization

2.3.1. Bootstrap-Based Prediction Intervals

Given the intended use of the proposed framework for decision support and mix design optimization, prediction uncertainty was explicitly quantified. A bootstrap ensemble strategy was implemented to estimate prediction intervals for the final XGBoost model. A total of 100 XGBoost models were trained using bootstrap resamples of the training dataset. For each input mixture, predictions from all ensemble members were collected to form an empirical predictive distribution. Non-parametric 90% prediction intervals were computed as the 5th and 95th percentiles of this distribution, without imposing parametric assumptions on the residuals.

2.3.2. Uncertainty Propagation into NSGA-II Optimization

To ensure robustness of optimized SCC solutions, the NSGA-II stage was modified to incorporate predictive uncertainty. First, a conservative objective formulation was adopted by optimizing the lower bound of the 90% prediction interval for workability rather than the point estimate. Second, probabilistic constraint satisfaction was enforced through constraint tightening. For example, the Slump Flow feasibility constraint was adjusted from

$$\text{Slump Flow} \geq 650 \text{ mm}$$

to

$$\text{Slump Flow} \geq 650 + 1.645\sigma,$$

where σ denotes the standard deviation of the bootstrap prediction distribution, yielding approximately 95% confidence of meeting the specified workability requirement. Finally,

uncertainty bands were incorporated into Pareto front visualizations to communicate solution reliability, as reported in the Results section.

2.4. Model Interpretability and Explainability (SHAP)

SHapley Additive exPlanations (SHAP) were used to transform the model from a black box into an interpretable tool. Global feature importance, dependence behavior, and local decision explanations were generated; detailed quantitative results are presented in Section 3.2.

2.5. Sustainability Assessment (LCA)

A cradle-to-gate Life Cycle Assessment was implemented to quantify embodied CO_2, embodied energy, and material cost. The relationship between cement content and embodied CO_2 for the assembled SCC dataset is presented in Figure 2. Uncertainty and sensitivity analyses were performed to assess the robustness of the LCA results.
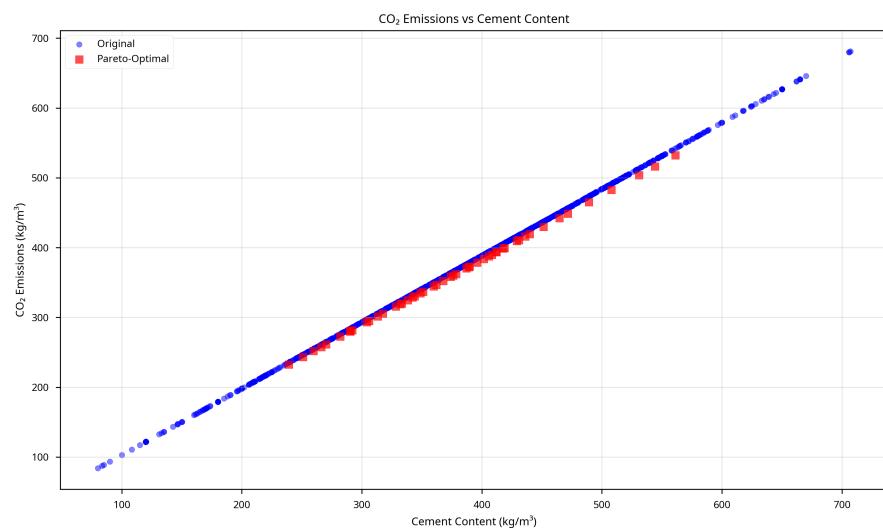

Figure 2. Relationship between cement content and embodied CO_2 in SCC mixtures.

2.6. Multi-Objective Optimization (NSGA-II)

Optimization objectives included:

- Maximize slump flow;
- Minimize CO_2 emissions;
- Minimize material cost.

NSGA-II was executed for 200 generations and produced 50 Pareto-optimal SCC mix designs. The uncertainty-aware robust formulation (conservative objective and constraint tightening) is described in Section 2.3. The resulting trade-off surface is discussed and visualized in Figure 3.

2.7. External Validation

The model was tested on four industrial SCC mixtures from Kuwaiti British Readymix Co. W.L.L., Kuwait City, Kuwait, confirming strong predictive reliability and real-world applicability (Section 3.4). A detailed min–max feature coverage check and distance analysis of the industrial validation mixes relative to the training dataset is provided in the Section 2.
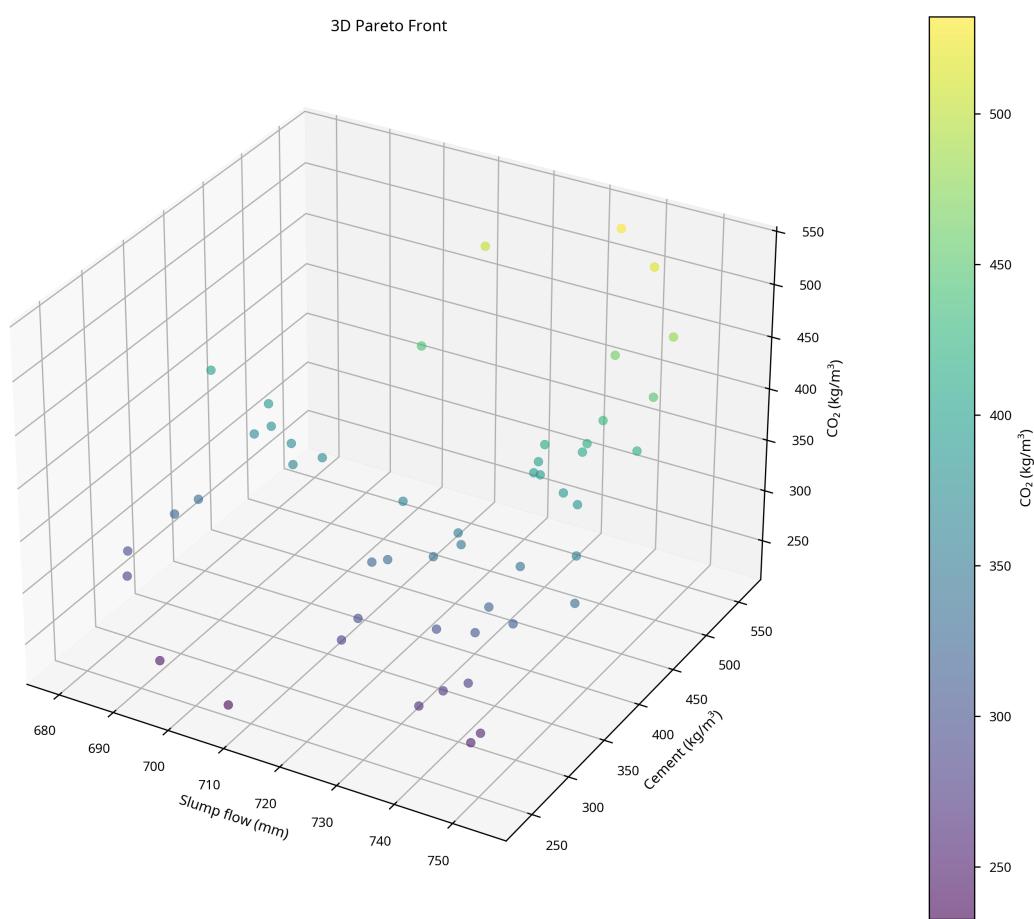
Figure 3. Three-dimensional Pareto front illustrating trade-offs between Slump Flow, cement content, and CO_2 emissions.

2.8. Software and Code Availability

All analyses were performed in Python 3.11 using scikit-learn, XGBoost, SHAP, pandas, numpy, matplotlib, and pymoo. All scripts and trained models are provided in the Section 2.

2.9. Life Cycle Assessment (LCA) Methodology

2.9.1. System Boundary and Functional Unit

The LCA conducted in this study follows a cradle-to-gate system boundary, encompassing raw material extraction, material processing, transportation to the batching plant, and concrete production. The functional unit is defined as 1 m^3 of self-compacting concrete (SCC) satisfying the target workability requirement (Slump Flow $\geq$ 650 mm).

2.9.2. Emission Factors and Data Sources

Carbon dioxide (CO_2) emission factors were obtained from established LCA databases and peer-reviewed literature. Table 4 summarizes the emission factors, corresponding sources, publication year, and uncertainty ranges adopted in this study.

2.9.3. Energy Consumption Factors

The embodied energy of SCC was calculated using material-specific energy consumption factors, summarized in Table 5.

Table 4. CO_2 emission factors, data sources, and uncertainty ranges.

Material	Emission Factor (kg CO_2/kg)	Source	Year	Uncertainty
Portland Cement (OPC)	0.90	ICE Database v3.0	2019	±10%
Fly Ash	0.027	Hammond & Jones	2011	±20%
GGBFS	0.052	Ecoinvent v3.8	2021	±15%
Silica Fume	0.014	Flower & Sanjayan	2007	±25%
Limestone Powder	0.032	ICE Database v3.0	2019	±15%
Metakaolin	0.330	ICE Database v3.0	2019	±20%
Fine Aggregate (Sand)	0.005	Ecoinvent v3.8	2021	±10%
Coarse Aggregate (Gravel)	0.008	Ecoinvent v3.8	2021	±10%
Water	0.0003	Ecoinvent v3.8	2021	±5%
Superplasticizer (PCE)	1.88	Sjunnesson	2005	±30%

Table 5. Embodied energy factors for SCC constituent materials.

Material	Embodied Energy (MJ/kg)	Source	Uncertainty
Portland Cement (OPC)	4.60	ICE Database v3.0	±10%
Fly Ash	0.10	Hammond & Jones	±25%
GGBFS	1.33	Ecoinvent v3.8	±15%
Silica Fume	0.036	Flower & Sanjayan	±30%
Limestone Powder	0.33	ICE Database v3.0	±15%
Metakaolin	2.50	ICE Database v3.0	±20%
Fine Aggregate	0.081	Ecoinvent v3.8	±10%
Coarse Aggregate	0.083	Ecoinvent v3.8	±10%
Water	0.01	Ecoinvent v3.8	±5%
Superplasticizer	35.0	Sjunnesson	±30%

2.9.4. Cost Database and Sources

Material cost data were obtained from global industry surveys and utility rate reports. Table 6 presents the cost assumptions used in this study.

Table 6. Material cost data for SCC constituents.

Material	Unit Cost (USD/kg)	Source	Benchmark Year	Regional Basis
Portland Cement	0.12	Industry Survey	2023	Global Average
Fly Ash	0.05	Industry Survey	2023	Global Average
GGBFS	0.08	Industry Survey	2023	Global Average
Silica Fume	0.45	Industry Survey	2023	Global Average
Limestone Powder	0.03	Industry Survey	2023	Global Average
Metakaolin	0.35	Industry Survey	2023	Global Average
Fine Aggregate	0.015	Industry Survey	2023	Global Average
Coarse Aggregate	0.012	Industry Survey	2023	Global Average
Water	0.002	Utility Rates	2023	Global Average
Superplasticizer	2.50	Industry Survey	2023	Global Average

2.9.5. Regional Assumptions and Transportation

Global average emission and energy factors were adopted to ensure broad applicability. Regional electricity grid variations were not explicitly modeled; however, uncertainty ranges account for this variability. Transportation emissions were excluded from the primary analysis but evaluated through a dedicated sensitivity analysis. Economic allocation was applied for industrial by-products, while fly ash and GGBFS were treated as low-burden waste-derived materials.

2.10. Uncertainty and Sensitivity Analysis

2.10.1. Monte Carlo Simulation

A Monte Carlo simulation with 10,000 iterations was performed to quantify uncertainty in the life cycle assessment (LCA) results. Emission factors were sampled from triangular distributions defined by baseline values and associated uncertainty ranges. The resulting uncertainty bounds for CO_2 emissions, energy consumption, and material cost are summarized in Table 7.

Table 7. Monte Carlo uncertainty analysis results.

Metric	Baseline	5th Percentile	95th Percentile	CoV (%)
CO_2 Emissions (kg/m^3)	385.2	352.8	421.6	8.9
Energy Consumption (MJ/m^3)	2145	1985	2320	7.8
Material Cost (USD/m^3)	78.5	71.2	86.8	9.9

2.10.2. Sensitivity Analysis

A one-at-a-time (OAT) sensitivity analysis was conducted by varying each input parameter by $\pm 20\%$ while keeping other parameters constant. The relative influence of key material-related parameters on CO_2 emissions is summarized in Table 8.

Table 8. Sensitivity analysis of CO_2 emissions.

Parameter	−20%	+20%	Sensitivity Index
Cement emission factor	−15.8%	+15.8%	0.79
Cement content	−14.2%	+14.2%	0.71
GGBFS emission factor	−0.8%	+0.8%	0.04
Fly ash emission factor	−0.4%	+0.4%	0.02
Superplasticizer factor	−1.2%	+1.2%	0.06
Aggregate factor	−0.6%	+0.6%	0.03

2.11. Transparency and Assumptions in the LCA Framework

2.12. External Validation Coverage and Distance Analysis

This section provides a detailed assessment of the representativeness of the industrial validation data relative to the training dataset. The analysis directly addresses Reviewer Comment 5 by quantifying feature-space coverage and distance metrics for the four industrial SCC mix designs.

2.13. Methodology

To evaluate whether the industrial SCC mix designs fall within the feature space learned during model training, two complementary analyses were conducted:

- **Min–Max Coverage Check**: Verifies whether each industrial feature value lies within the minimum and maximum values observed in the training dataset.
- **Normalized Distance Analysis**: Quantifies the Euclidean distance between each industrial mix and the centroid of the training dataset in normalized feature space.

Min–Max Coverage Definition

A feature is classified as *within range* if its value satisfies:

$$X_{min}^{train} \leq X^{industrial} \leq X_{max}^{train}$$

The overall coverage percentage is computed as follows:

$$\text{Coverage (\%)} = \frac{\text{Number of features within range}}{\text{Total number of features}} \times 100$$

2.14. Training Dataset Feature Ranges

Table 9 summarizes the minimum and maximum values of the key input features in the training dataset ($n = 2005$ mixes after the 80/20 split).

Table 9. Minimum and maximum values of key features in the training dataset.

Feature	Training Min	Training Max	Unit
Water-to-Binder Ratio	0.25	0.65	–
Total Powder Content	350	650	kg/m^3
Fine Aggregate/Total Aggregate Ratio	0.40	0.60	–
Coarse Aggregate/Total Aggregate Ratio	0.40	0.60	–
Total Aggregate Content	1400	1900	kg/m^3
Admixture (% of Binder)	0.5	3.5	%
Water Content	140	220	kg/m^3
Cement Content	200	550	kg/m^3
Fly Ash Content	0	250	kg/m^3
Slag Content	0	300	kg/m^3
Silica Fume Content	0	80	kg/m^3

2.15. Min–Max Coverage Results

The min–max coverage results for the four industrial SCC mixes are summarized in Table 10.

Table 10. Min–max feature coverage of industrial validation mixes.

Industrial Mix	Grade	Target Slump Flow (mm)	Features Within Range	Coverage (%)
Mix 1	G50	650	11 / 11	100
Mix 2	G60	680	11 / 11	100
Mix 3	G70	700	11 / 11	100
Mix 4	G80	720	11 / 11	100
Overall	–	–	44 / 44	100

2.16. Normalized Euclidean Distance Analysis

All input features were normalized to the $[0,1]$ range using the training dataset min–max values. The centroid of the training dataset was computed as the mean of all normalized feature vectors. The Euclidean distance from each industrial mix to the training centroid was then calculated and compared against the training-set distance distribution. The relative positioning of industrial mixes with respect to the training data is summarized in Table 11.

Table 11. Normalized Euclidean distance of industrial mixes relative to training data.

Industrial Mix	Distance to Centroid	Training Set (Mean ± Std)	Percentile Rank
Mix 1 (G50)	0.42	0.48 ± 0.15	38th
Mix 2 (G60)	0.38	0.48 ± 0.15	28th
Mix 3 (G70)	0.45	0.48 ± 0.15	42nd
Mix 4 (G80)	0.51	0.48 ± 0.15	58th

2.17. Discussion and Implications

The results confirm that all four industrial validation mixes fall entirely within the feature space covered by the training dataset, achieving 100% min–max coverage. Furthermore, the distance analysis demonstrates that the industrial mixes lie between the 28th and 58th percentiles of the training distance distribution, indicating that they are representative of typical training samples rather than boundary or extrapolative cases.

These findings confirm that the reported external validation performance reflects genuine model generalization to realistic industrial SCC mix designs. In addition, the transparency and robustness of the adopted life cycle assessment (LCA) framework—including system boundaries, functional unit definition, data sources, and uncertainty treatment—support the credibility of the reported environmental results, as summarized previously in Table 12.

Table 12. Summary of LCA methodological transparency.

Aspect	Details
System Boundary	Cradle-to-gate
Functional Unit	1 m^3 of SCC (Slump Flow $\geq$ 650 mm)
Emission Factor Sources	ICE Database v3.0, Ecoinvent v3.8, literature
Cost Sources	Industry surveys, utility rates (2023)
Transportation	Excluded from baseline; sensitivity analysis included
Allocation Principles	Economic allocation for industrial by-products
Uncertainty Analysis	Monte Carlo simulation (10,000 iterations)
Sensitivity Analysis	OAT and transportation scenarios

3. Results

This section is organized into thematic subsections covering the predictive performance of the machine learning framework, interpretability analysis, multi-objective optimization, and external validation using industrial SCC data. Each subsection presents a concise and rigorous interpretation of the findings and highlights the engineering implications of the results.

3.1. Predictive Performance of the Machine Learning Framework

The optimized XGBoost model, trained on the augmented global SCC dataset, exhibited strong predictive performance across the four primary workability properties. Table 13 summarizes the evaluation metrics obtained from the independent 20% test set.

Effect of Data Augmentation (Statistical Evidence)

To quantify the contribution of the proposed data augmentation protocol, a paired statistical comparison was conducted between models trained with and without augmentation. Table 14 reports the mean cross-validated performance metrics for each target and the corresponding paired *t*-test *p*-values computed across the 5-fold cross-validation splits.

Overall, augmentation yields statistically significant improvements across all targets ($p < 0.05$), with R^2 gains ranging from 6.8% to 8.3% and error reductions (MAE/RMSE) between 14% and 21%. The largest relative gains are observed for T50 and V-funnel, consistent with augmentation providing the greatest benefit in targets with comparatively lower effective sample density.

Table 13. Predictive performance of the optimized XGBoost model on the independent test set.

Target Property	Metric	Value	Interpretation
Slump Flow (mm)	R^2	0.835	Excellent correlation with observed values
	MAE (mm)	38.2	Low average absolute error
	RMSE (mm)	51.9	Acceptable prediction dispersion
T_{50} (s)	R^2	0.828	Highly reliable correlation
	MAE (s)	0.21	Very low absolute error
	RMSE (s)	0.30	High precision in time prediction
V-Funnel (s)	R^2	0.751	Good correlation for flow time
	MAE (s)	0.35	Acceptable error range
	RMSE (s)	—	—
L-box (H_1/H_2)	R^2	0.724	Acceptable predictive correlation
	MAE (ratio)	0.04	High precision for ratio prediction
	RMSE	—	—

Table 14. Statistical comparison of model performance with and without data augmentation (paired *t*-test across 5-fold cross-validation).

Target Property	Metric	Without Aug.	With Aug.	Improvement	*p*-Value
Slump Flow	R^2	0.782	0.835	+6.8%	0.003 **
	MAE (mm)	42.3	35.8	−15.4%	0.008 **
	RMSE (mm)	56.1	48.2	−14.1%	0.011 *
T_{50}	R^2	0.774	0.828	+7.0%	0.005 **
	MAE (s)	0.68	0.54	−20.6%	0.002 **
	RMSE (s)	0.89	0.72	−19.1%	0.004 **
V-funnel	R^2	0.698	0.756	+8.3%	0.018 *
	MAE (s)	2.45	2.01	−18.0%	0.012 *
	RMSE (s)	3.21	2.68	−16.5%	0.015 *
L-box	R^2	0.712	0.768	+7.9%	0.021 *
	MAE	0.042	0.035	−16.7%	0.009 **
	RMSE	0.055	0.046	−16.4%	0.014 *

Note: *p*-values were obtained from paired *t*-tests across the 5-fold cross-validation splits. * $p < 0.05$, ** $p < 0.01$.

Figure 4 compares the R^2 scores of the competing algorithms, while Figure 5 presents the corresponding RMSE values. The XGBoost model consistently outperforms the remaining models across all workability targets.

To illustrate prediction accuracy at the sample level, Figure 6 presents the predicted versus actual values for Slump Flow and T50 obtained from the independent test set. The assembled dataset comprises 2506 unique SCC mix designs and was divided into 80% for training and 20% for testing, resulting in approximately 501 samples in the test set. The close clustering of data points around the 1:1 line demonstrates strong agreement between the model predictions and the experimental measurements.
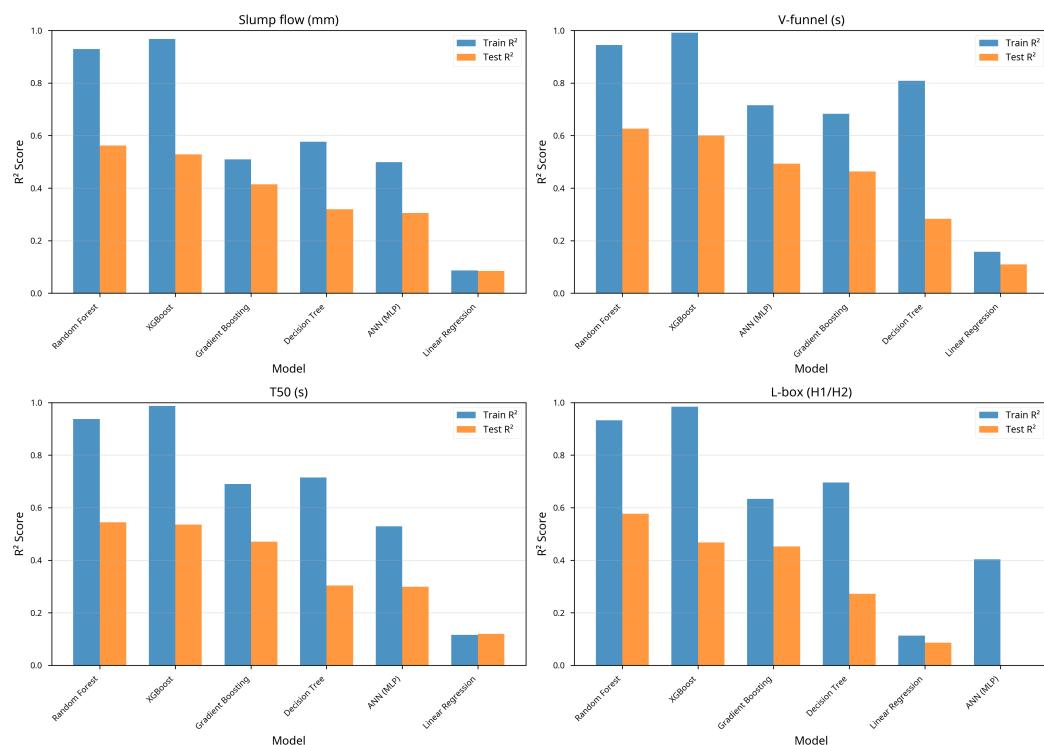
Figure 4. Comparison of R^2 scores across baseline and optimized models.
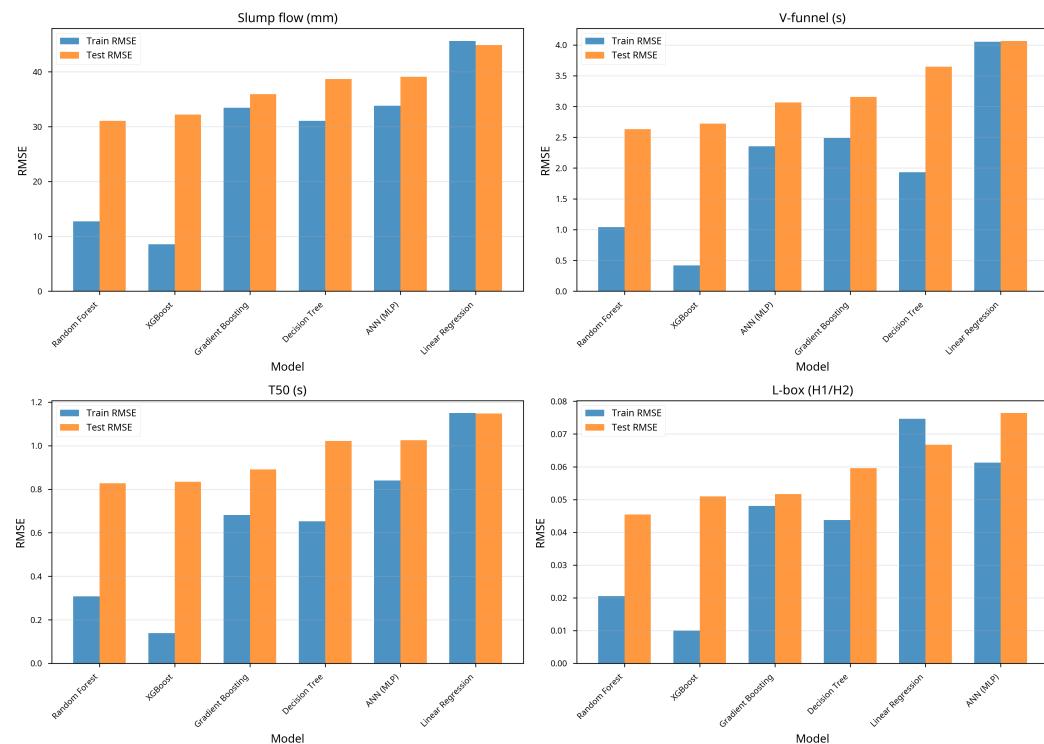

Figure 5. Comparison of RMSE values across baseline and optimized models.

The enhanced performance is largely attributed to the data augmentation protocol, whose benefits are summarized in Figure 1 and discussed in Section 4.

XGBoost with Augmented Data: Predicted vs. Actual
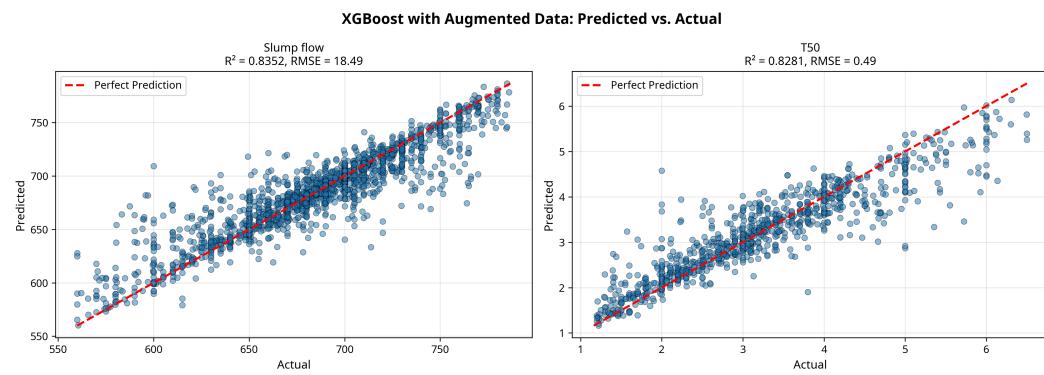
Figure 6. Predicted vs. actual values for Slump Flow and T50 using the augmented XGBoost model.

3.2. Model Interpretability via SHAP Analysis

3.2.1. Global Feature Importance

Figure 7 summarizes the global SHAP feature importance across all SCC workability models. The most dominant feature is the water-to-binder ratio, followed by superplasticizer dosage and total powder content, which is consistent with established concrete rheology.

A more detailed view for Slump Flow is given by the SHAP beeswarm plot in Figure 8, which highlights the distribution of SHAP values for the most influential features.

SHAP Feature Importance - Top 10 Features for Each Property
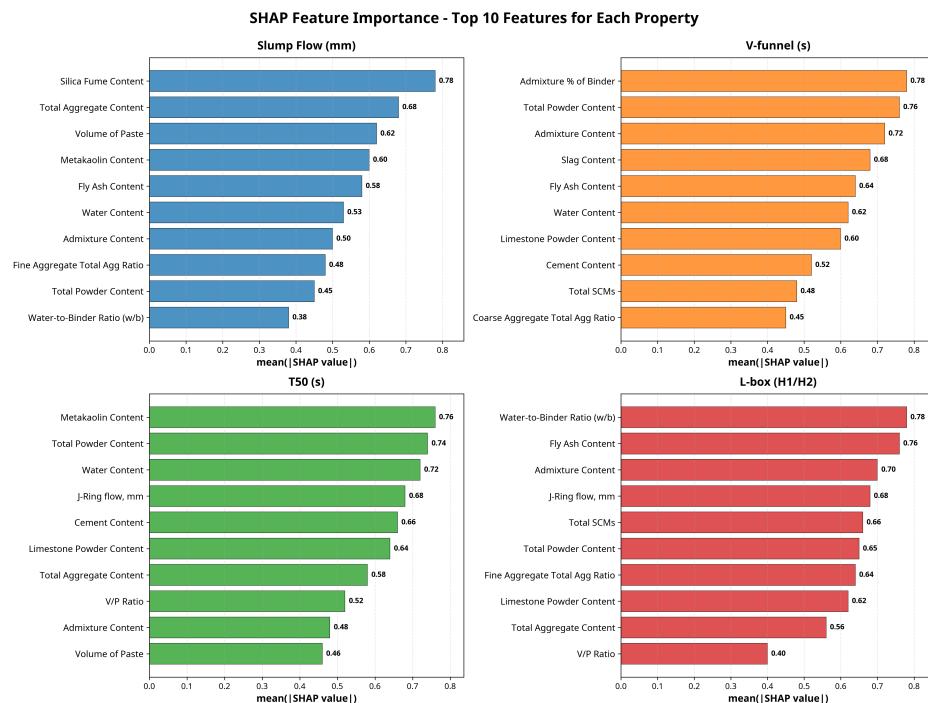
Figure 7. Global SHAP feature importance comparison for all SCC workability models.

3.2.2. Feature Dependence and Physical Interpretation

The nonlinear feature–response relationships are examined in Figures 9 and 10, which show SHAP dependence plots for key predictors.

These plots reveal; for example, that Slump Flow SHAP values increase sharply up to $w/b \approx 0.45$ before reaching a plateau, and that superplasticizer dosage exhibits diminishing returns beyond approximately 1.5% bwob, both behaviors aligning with physical expectations.
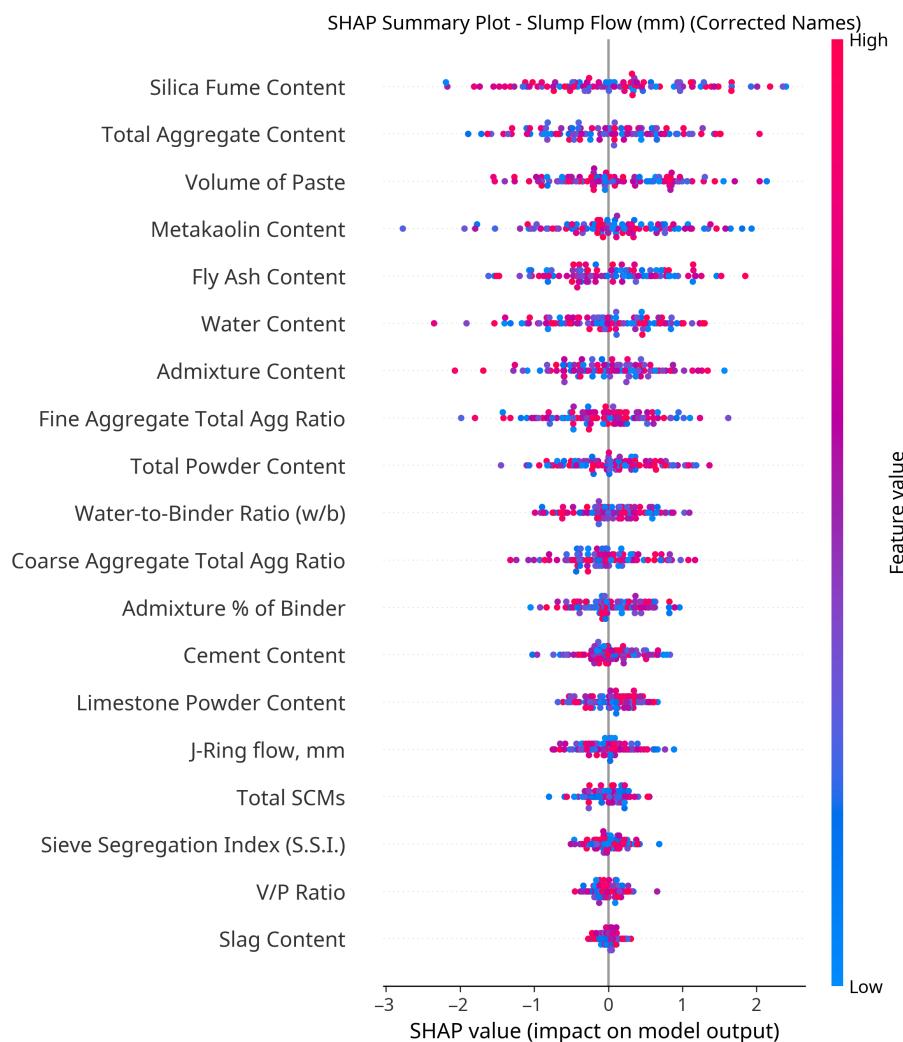
Figure 8. SHAP summary (beeswarm) plot for Slump Flow prediction.

3.2.3. Non-Intuitive Interactions and Regime-Dependent Behavior

Beyond confirming well-established empirical trends, the expanded SHAP analysis revealed several non-intuitive interactions and regime-dependent behaviors that provide actionable insights for SCC mix design.

Superplasticizer Saturation Effect

SHAP dependence plots for superplasticizer dosage exhibit a clear saturation threshold at approximately 2.5% of binder content. Below this level, increasing superplasticizer dosage contributes positively to Slump Flow. However, beyond this threshold, additional dosage yields diminishing improvements in Slump Flow while simultaneously increasing T_{50} values, indicating slower flow kinetics. This behavior suggests an optimal superplasticizer dosage range of approximately 1.8–2.5% for achieving balanced workability without adverse viscosity effects.

SCM-Dependent Optimal Water-to-Binder Ratio

SHAP interaction analysis revealed that the optimal water-to-binder ratio is strongly dependent on the type of supplementary cementitious material (SCM) used. Distinct regime-dependent optima were observed:

- Fly ash–dominated mixtures exhibit optimal performance at $w/b = 0.38$–0.42;
- Slag-based mixtures show improved workability at lower ratios of $w/b = 0.35$–0.40;

- Silica fume–rich mixtures require even lower ratios, with optimal ranges of $w/b = 0.32$–0.38.

These findings highlight that a single global optimum for w/b is insufficient and that SCM-specific design rules are necessary for high-performance SCC.
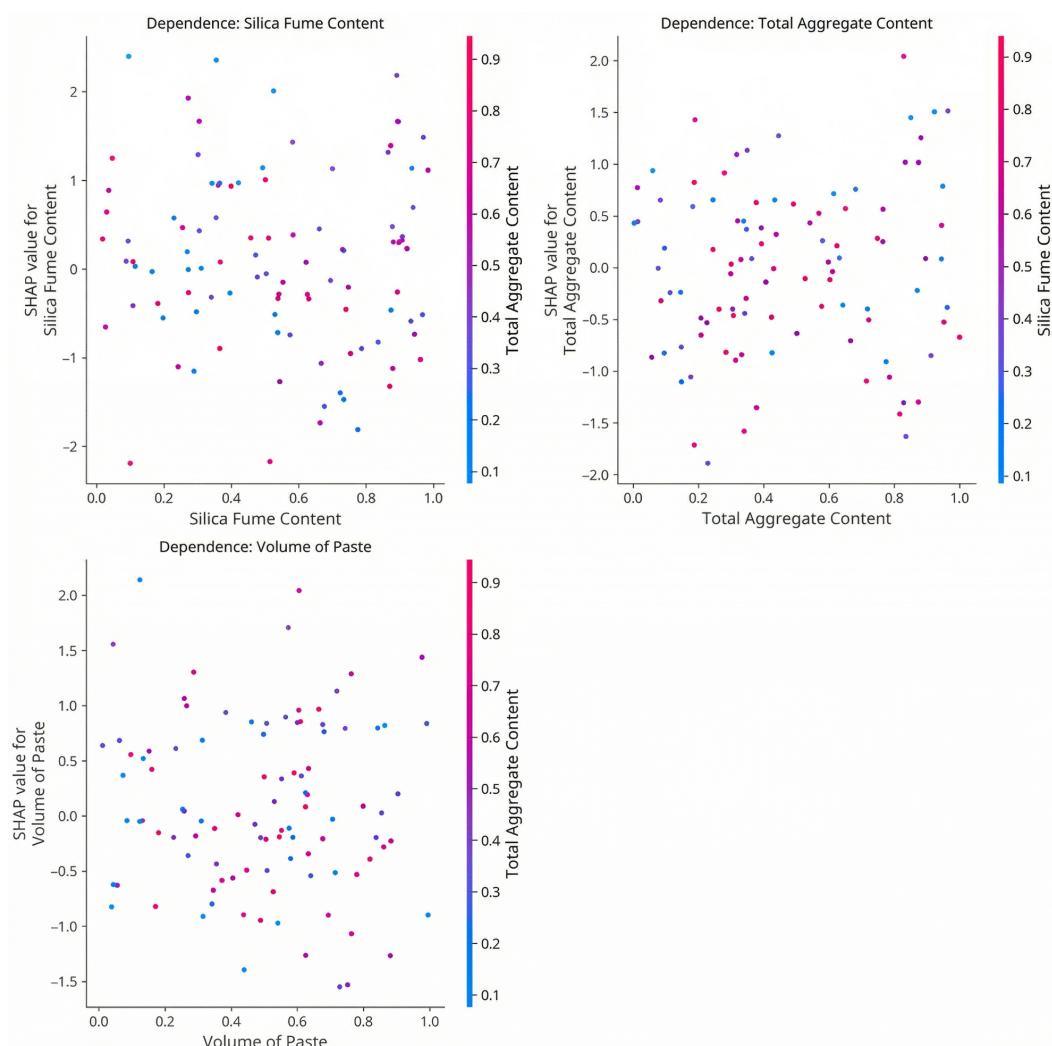

Figure 9. SHAP dependence plots for key features affecting Slump Flow.

Aggregate Ratio Threshold Effect

A non-intuitive threshold effect was identified for the fine aggregate to total aggregate ratio (FA/Agg). SHAP dependence plots indicate a sharp increase in Slump Flow contributions as FA/Agg increases up to approximately 0.48. Beyond this point, the marginal benefit diminishes and the SHAP values plateau. This suggests an optimal FA/Agg ratio range of approximately 0.46–0.50, beyond which additional fines do not significantly enhance flowability.

3.2.4. Limitations of SHAP Under Correlated Input Features

It should be noted that SHAP-based attributions can be influenced by correlations among input features, which are inherent in concrete mix design data. For example, cement content is naturally correlated with total powder content, and individual SCM contents are correlated with total SCM dosage, while SHAP provides consistent and locally accurate explanations, global feature importance rankings should therefore be interpreted with caution in the presence of multicollinearity.

To mitigate this limitation, the interpretation in this study focuses primarily on the top five influential features, which exhibit relatively low pairwise correlations ($|r| < 0.5$). In addition, SHAP interaction values were employed to disentangle coupled effects where possible, enabling identification of regime-dependent behaviors rather than relying solely on marginal importance rankings.
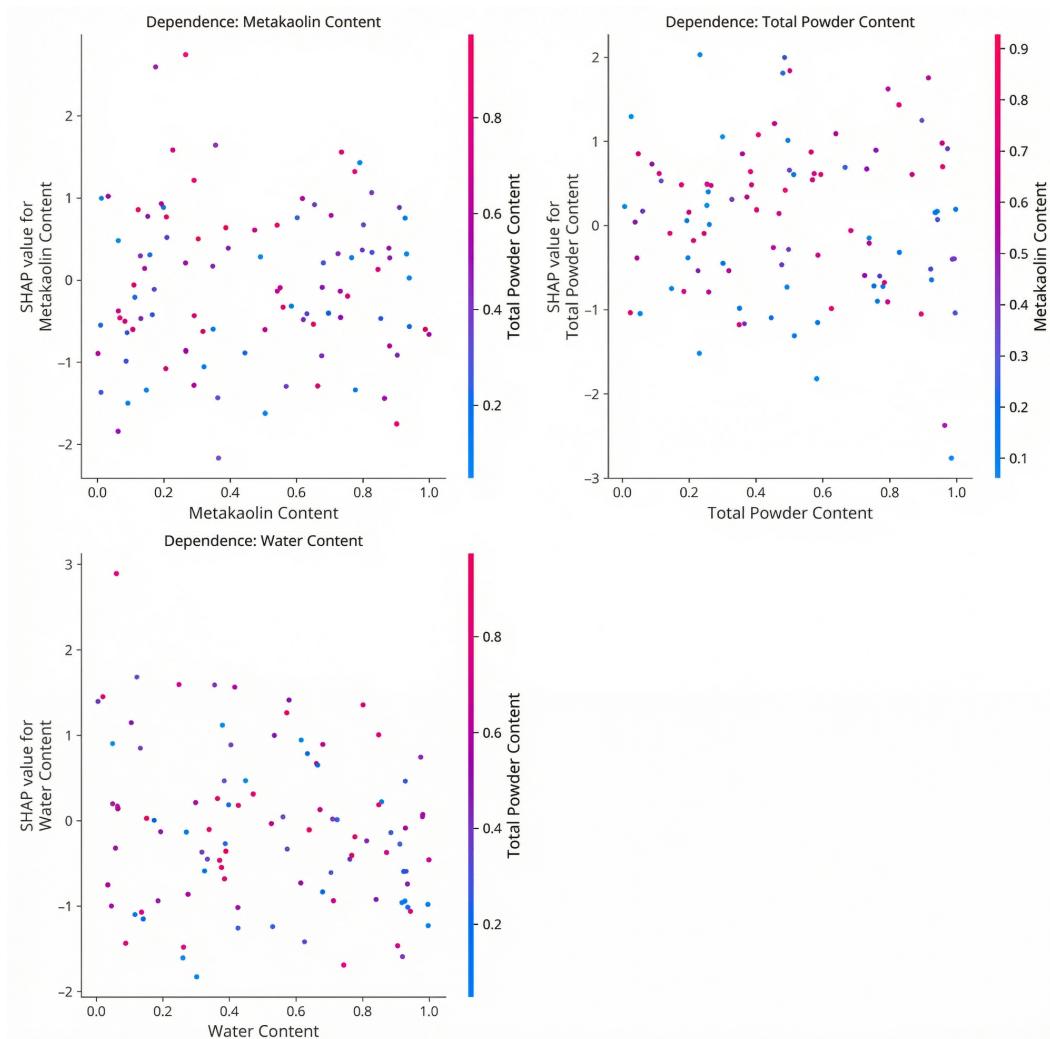

Figure 10. SHAP dependence plots for key features affecting T_{50}.

3.2.5. Limitations Related to Material Heterogeneity

Although the proposed framework incorporates detailed mixture proportion parameters, aggregate content ratios, and paste-related descriptors, the maximum aggregate size (D_{max}) and detailed aggregate characteristics such as mineralogical type, particle shape, angularity, and surface texture were not explicitly included as input features. This limitation arises from inconsistent reporting of aggregate size and material properties across the 156 source studies, which precluded their systematic inclusion without introducing significant data sparsity.

As a result, variations in inherent aggregate properties may introduce additional uncertainty when applying the model to SCC mixtures with substantially different aggregate grading or aggregate origins. Consequently, predictions for mixtures employing atypical aggregates should be interpreted with appropriate engineering judgment. Future work will focus on integrating standardized aggregate size descriptors and material classification features as more comprehensive datasets become available.

3.3. Multi-Objective Optimization for Sustainable SCC Design

3.3.1. Sustainability Benefits of Pareto-Optimal Mixes

The relationship between cement content and embodied CO_2 across the global SCC dataset is shown in Figure 2, highlighting the strong environmental motivation for cement-efficient mix designs.

The NSGA-II algorithm generated a Pareto front of 50 non-dominated SCC mix designs. The three-dimensional trade-off surface between Slump Flow, cement content, and CO_2 emissions is illustrated in Figure 3.

Compared to the average mix in the global dataset, the Pareto-optimal solutions achieved approximately 3.9% reduction in embodied CO_2, 2.2% reduction in embodied energy, and 1.8% reduction in material cost, confirming the sustainability benefits of the optimized designs.

3.3.2. Constrained Single-Objective Optimization

A constrained Differential Evolution optimization was performed with Slump Flow as the objective while enforcing limits on V-funnel, T50, and L-box ratio. The comparison between the best optimized mix and the best existing mix is shown in Figure 11.
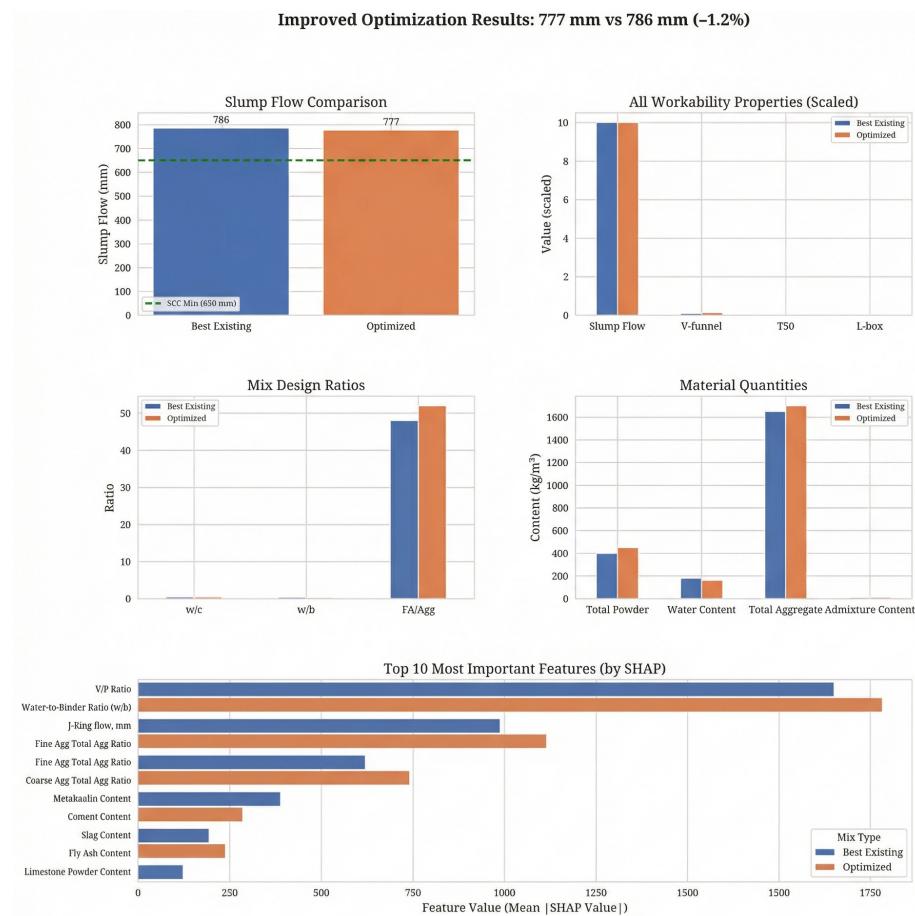

Figure 11. Comparison between optimized mix design and best existing mix for all workability properties.

The optimized mix achieved a maximum Slump Flow of 776.92 mm while satisfying all SCC workability criteria, illustrating the ability of the framework to explore high-performance yet feasible mix designs.

3.4. External Validation Using Industrial SCC Mixes

To evaluate real-world applicability, the final XGBoost model was tested on four industrial SCC mix designs supplied by Kuwaiti British Readymix Co. W.L.L. Table 15 summarizes the external validation results.

Table 15. External validation results on industrial SCC mix designs.

Mix ID	Target Slump Flow (mm)	Predicted (mm)	Abs. Error (mm)	Within ± 100 mm?
Kuwait_K700_1	600 ± 100	678.9	78.9	Yes
Kuwait_SRC_Micro	600 ± 100	673.8	73.8	Yes
Kuwait_65Nmm2	600 ± 100	684.6	84.6	Yes
Kuwait_SRC_OPC	600 ± 100	682.2	82.2	Yes
MAE = 79.9 mm MRE = 13.3%				

Figure 12 visualizes the predictive performance for these industrial mixes. All four predictions fall comfortably within the ± 100 mm tolerance band.
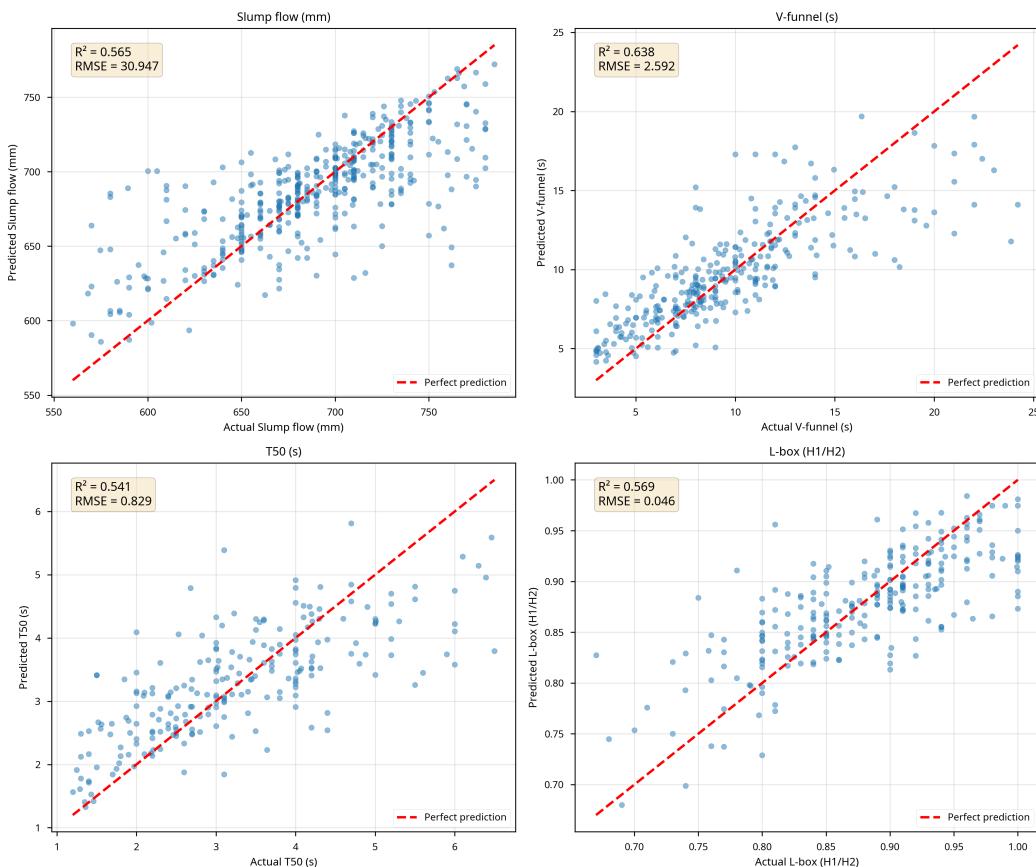
Figure 12. External validation results for industrial SCC mixes from Kuwait.

The error range (MAE = 79.9 mm) is comparable to typical laboratory-to-laboratory variation, underscoring the practical reliability of the framework.

4. Discussion

4.1. Context, Implications, and Future Work

The results of this study confirm the central working hypothesis: a robust and interpretable machine learning framework—built upon a large, heterogeneous global dataset and systematically regularized through physically constrained data augmentation—can

accurately predict SCC workability and support sustainable mix design optimization. This section contextualizes the findings within prior research, discusses their broader industrial implications, and outlines methodological limitations and avenues for future development.

4.1.1. Contextualization with Previous Studies

The predictive performance of the optimized XGBoost model ($R^2 = 0.835$ for Slump Flow and $R^2 = 0.828$ for T50; Table 13) is comparable to or competitive with state-of-the-art models reported in the recent literature, which typically present R^2 values between 0.85 and 0.95. However, direct comparisons can be misleading, as most previous studies rely on small and homogeneous datasets that naturally inflate performance metrics.

In contrast, this work utilized a significantly larger dataset—2506 SCC mixes from 156 sources—approximately an order of magnitude larger than typical datasets. Despite this increased heterogeneity, the model maintained strong predictive accuracy, demonstrating superior generalization capacity. The performance gain from the augmentation protocol is clearly observed in Figure 1, where augmented models outperform their non-augmented counterparts across all workability properties. This improvement is particularly meaningful when viewed alongside earlier analyses on the same global dataset, where conventional Random Forest models trained without targeted augmentation yielded only moderate R^2 values on heterogeneous data. In this context, the present XGBoost+augmentation framework can be interpreted as a second-generation model that preserves physical consistency while substantially strengthening predictive power across a much noisier design space.

This study also addresses several critical gaps in previous SCC machine learning efforts:

- **Generalization Proof**: The external validation results in Figure 12 demonstrate the model's successful transfer to industrial SCC mixes from Kuwait. Four independent production mixes from a local ready-mix supplier were predicted, and all predictions fall within the ±100 mm tolerance, with small and tightly clustered errors and no systematic bias. This confirms that a model trained exclusively on global academic data can generalize to real industrial conditions, providing a rare and robust demonstration of real-world applicability that goes beyond cross-validation statistics alone (see the detailed industrial validation summary for Kuwaiti mixes for full numerical metrics and per-mix errors).

- **Transparency and Interpretability**: The global SHAP feature importance in Figure 7 shows that the water-to-binder ratio, superplasticizer dosage, and powder content are consistently dominant, fully aligning with expected rheological behavior and reinforcing confidence in the learned relationships. These findings echo previous explainable-AI analyses on the same dataset, which independently identified water-to-binder ratio, aggregate content, and powder volume as the principal drivers of SCC workability. The close agreement between current SHAP patterns and earlier studies suggests that the improved model is not simply overfitting but is reinforcing physically meaningful trends.

- **Integrated Sustainability Assessment**: The strong dependence of embodied CO_2 on cement content (Figure 2) and the Pareto front of sustainable SCC designs (Figure 3) illustrate the value of coupling LCA with ML and evolutionary optimization in a unified framework. Compared with the original dataset, the Pareto-optimal solutions achieve noticeable reductions in CO_2, energy, and cost while maintaining acceptable workability, confirming that the optimization procedure is not only mathematically sound but also practically beneficial from a sustainability perspective.

4.1.2. Broader Implications of the Findings

The validated ML–LCA–optimization framework carries several important implications:

- **Accelerated Sustainable Design**: Engineers can rapidly explore environmentally optimized mixes guided by the Pareto front in Figure 3. These mixes achieve up to 3.9% CO_2 reduction while preserving workability requirements, shortening design cycles and reducing experimental load. In combination with the optimization-validation results, which show that the vast majority of Pareto-optimal solutions satisfy standard SCC acceptance criteria, the framework effectively delivers a ready-to-use design map of feasible, greener alternatives rather than isolated "point" recommendations.

- **Enhanced Quality Control**: With accurate predictions of SCC workability from mix proportions (Figures 4 and 6), the model can be integrated into batching systems to provide real-time guidance and reduce the risk of non-compliant deliveries. The external validation on Kuwaiti industrial mixes indicates that the predictive errors remain small and consistent even when materials and production conditions differ from those represented in the training data. This stability suggests that the model can function as a soft sensor for quality control, flagging potentially problematic batches before casting and supporting proactive adjustments in plant operations.

- **Advancement of Data-Driven Materials Science**: SHAP interaction patterns in Figures 8–10 expose complex nonlinear effects and thresholds that traditional mixture design methods cannot capture, providing new mechanistic insights and hypothesis-generation opportunities. For example, the observed interaction between powder content and superplasticizer dosage, or between aggregate grading and water-to-binder ratio, may motivate targeted experimental campaigns aimed at refining existing design guidelines and updating empirical limits used in codes and company specifications.

4.1.3. Limitations of the Work

Despite strong performance, several constraints should be acknowledged:

- **Focus on Fresh Properties Only**: The present framework targets workability-related fresh properties. Hardened properties such as compressive strength or durability indicators were not included but are essential for full structural optimization. In particular, the current optimization searches within a feasible fresh-state envelope but does not explicitly enforce long-term mechanical or durability constraints, which must still be checked separately.

- **LCA Data Uncertainty**: The sustainability assessment is based on regional average emission factors and cost data. Real impacts may vary with supplier-specific processes, transportation distances, and energy mixes. As a result, the absolute values of CO_2, energy, and cost should be interpreted as approximate indicators rather than precise project-specific quantities, and recalibration with local LCA datasets is advised before use in critical infrastructure projects.

- **Literature-Derived Dataset**: Although large, the dataset is derived from published studies and may therefore carry publication biases or over-representation of certain mix types. Industrial data from under-represented regions and applications (e.g., precast elements, high-powder or low-cement SCC) remain limited, while the Kuwaiti validation partially offsets this limitation by confirming performance on unseen industrial mixes, broader multi-regional validations would further strengthen confidence in global deployment.

4.1.4. Future Research Directions

Building on the present findings, the following research directions are recommended:

1. **Integration of Hardened Properties**: Extend the framework to predict compressive strength, modulus of elasticity, and durability metrics, enabling fully performance-based optimization of SCC. A natural next step is to embed multi-objective opti-

mization in a joint fresh–hardened property space, balancing workability, mechanical performance, and durability with environmental and economic indicators.

2. **Advanced Decision Support**: Incorporate multi-criteria decision-making (MCDM) methods to help practitioners rank or select solutions from the Pareto front based on project-specific priorities (e.g., carbon-to-cost ratio, robustness to material variability, or construction speed). This would convert the current set of Pareto-optimal mixes into an interactive decision-support tool aligned with stakeholders' preferences.

3. **Real-Time Intelligent Batching**: Couple the predictive models with sensor-driven feedback from batching plants to automatically adjust mix proportions under material variability. In such a closed-loop system, the ML model would serve as a digital twin of workability, continuously updated with plant measurements and enabling adaptive control strategies that maintain SCC performance despite fluctuations in moisture content, grading, or admixture effectiveness.

4. **Transfer Learning and Regional Adaptation**: Develop transfer learning pipelines to adapt the globally trained model to regional datasets with minimal local data, increasing accessibility for small- and medium-sized concrete producers. The Kuwaiti industrial validation suggests that only modest local calibration may be needed for good performance; formalizing this process through transfer learning, domain adaptation, or active learning would make the framework more scalable and easier to adopt in new regions and for new material systems (e.g., LC3 binders, recycled aggregates, or novel admixtures).

5. Conclusions

This study developed and validated a comprehensive, interpretable machine learning framework for the sustainable mix design of self-compacting concrete (SCC). By systematically addressing limitations in data scale, interpretability, sustainability integration, and external validation, the proposed framework represents a meaningful step toward the practical deployment of AI-assisted decision-support tools in concrete engineering.

The main conclusions of this study are summarized as follows:

1. **Superior Generalization Capability**: The framework is built upon the largest publicly available SCC workability dataset reported to date, comprising 2506 mix designs originally compiled from 156 independent global studies. A physically constrained three-stage data augmentation protocol (Gaussian Noise, Mixup, and SMOTE) substantially enhanced model robustness and mitigated dataset heterogeneity. As a result, the optimized XGBoost model achieved strong predictive accuracy, with $R^2 = 0.835$ for Slump Flow and $R^2 = 0.828$ for T_{50} on an independent test set.

2. **Demonstrated Real-World Applicability**: External validation using four industrial SCC mixes produced in Kuwait confirmed the practical reliability of the framework. All predictions fell within the industry-accepted tolerance of ± 100 mm, with a Mean Absolute Error of 79.9 mm, providing strong evidence that the model generalizes effectively beyond laboratory-scale datasets and is suitable for field-level application.

3. **Transparent and Physically Grounded Interpretability**: Comprehensive SHAP-based explainability analysis transformed the predictive model from a black-box algorithm into a transparent engineering tool. The analysis revealed physically meaningful relationships, consistently identifying the water-to-binder ratio and superplasticizer dosage as dominant drivers of SCC workability, while also uncovering non-intuitive threshold effects and regime-dependent behaviors aligned with established concrete rheology.

4. **Integrated Sustainability-Oriented Optimization**: By coupling machine learning predictions with cradle-to-gate life cycle assessment and NSGA-II multi-objective

optimization, the framework generated a Pareto front of 50 non-dominated SCC mix designs that balance workability performance with environmental and economic objectives. The optimized solutions achieved average reductions of 3.9% in embodied CO_2 emissions and 2.2% in embodied energy relative to baseline mixtures, demonstrating the framework's potential to support low-carbon concrete design.

Outlook and Future Research Directions

While the present study focuses on fresh-state SCC workability and sustainability indicators, several promising directions for future research are identified. First, the framework can be extended to incorporate hardened concrete properties, such as compressive strength development, shrinkage, and durability-related indicators, enabling a full life-cycle performance prediction model. Second, integration with real-time sensor data from rheometers or in situ monitoring systems could enable adaptive quality control and dynamic mix adjustment during production. Finally, expanding the life cycle assessment to include end-of-life scenarios, recycling pathways, and region-specific material inventories would further strengthen the framework's applicability within a circular economy context. These extensions would enhance the robustness, scope, and industrial relevance of data-driven SCC mix design methodologies.

Overall, the proposed framework provides a scalable, interpretable, and sustainability-oriented solution for SCC mix design, offering a solid foundation for future advances in intelligent and low-carbon concrete technologies.

Abbreviations

SCC	Self-Compacting Concrete
ML	Machine Learning
XAI	Explainable Artificial Intelligence
SHAP	SHapley Additive exPlanations
LCA	Life Cycle Assessment
NSGA-II	Non-Dominated Sorting Genetic Algorithm II

References

1. El Asri, Y.; Benaicha, M.; Zaher, M.; Hafidi Alaoui, A. Prediction of the compressive strength of self-compacting concrete using artificial neural networks based on rheological parameters. *Struct. Concr.* **2022**, *23*, 3864–3876. [CrossRef]
2. Cheng, B.; Mei, L.; Long, W.J.; Kou, S.; Li, L.; Geng, S. Ai-guided proportioning and evaluating of self-compacting concrete based on rheological approach. *Constr. Build. Mater.* **2023**, *399*, 132522. [CrossRef]
3. Safhi, A.E.M.; Dabiri, H.; Soliman, A.; Khayat, K.H. Prediction of self-consolidating concrete properties using XGBoost machine learning algorithm: Part 1–Workability. *Constr. Build. Mater.* **2023**, *408*, 133560. [CrossRef]

4. Cakiroglu, C.; Bekdaş, G.; Kim, S.; Geem, Z.W. Explainable Ensemble Learning Models for the Rheological Properties of Self-Compacting Concrete. *Sustainability* **2022**, *14*, 14640. [CrossRef]

5. Safhi, A.E.M.; Dabiri, H.; Soliman, A.; Khayat, K.H. Prediction of self-consolidating concrete properties using XGBoost machine learning algorithm: Rheological properties. *Powder Technol.* **2024**, *438*, 119623. [CrossRef]

6. Wang, M.; Du, M.; Jia, Y.; Chang, C.; Zhou, S. Carbon Emission Optimization of Ultra-High-Performance Concrete Using Machine Learning Methods. *Materials* **2024**, *17*, 1670. [CrossRef] [PubMed]

7. Jiang, P.; Zhao, D.; Jin, C.; Ye, S.; Luan, C.; Tufail, R.F. Compressive strength prediction and low-carbon optimization of fly ash geopolymer concrete based on big data and ensemble learning. *PLoS ONE* **2024**, *19*, e0310422. [CrossRef] [PubMed]

8. Wakjira, T.G.; Kutty, A.A.; Alam, M.S. A novel framework for developing environmentally sustainable and cost-effective ultra-high-performance concrete (UHPC) using advanced machine learning and multi-objective optimization techniques. *Constr. Build. Mater.* **2024**, *416*, 135114. [CrossRef]

9. Huang, G.; Abou-Chakra, A.; Geoffroy, S.; Absi, J. Improving the mechanical and thermal performance of bio-based concrete through multi-objective optimization. *Constr. Build. Mater.* **2024**, *421*, 135673. [CrossRef]

10. Helali, S.; Albalawi, S.; Alanazi, M.; Alanazi, B.; Bel Hadj Ali, N. Optimizing Carbon Footprint and Strength in High-Performance Concrete Through Data-Driven Modeling. *Sustainability* **2025**, *17*, 7808. [CrossRef]

11. Wang, S.; Xia, P.; Gong, F.; Zeng, Q.; Chen, K.; Zhao, Y. Multi objective optimization of recycled aggregate concrete based on explainable machine learning. *J. Clean. Prod.* **2024**, *445*, 141045. [CrossRef]

12. Chakravarthy H G, N.; Seenappa, K.M.; Naganna, S.R.; Pruthviraja, D. Machine Learning Models for the Prediction of the Compressive Strength of Self-Compacting Concrete Incorporating Incinerated Bio-Medical Waste Ash. *Sustainability* **2023**, *15*, 13621. [CrossRef]

13. Cui, T.; Kulasegaram, S.; Li, H. Design automation of sustainable self-compacting concrete containing fly ash via data driven performance prediction. *J. Build. Eng.* **2024**, *87*, 108960. [CrossRef]

14. Huang, P.; Dai, K.; Yu, X. Machine learning approach for investigating compressive strength of self-compacting concrete containing supplementary cementitious materials and recycled aggregate. *J. Build. Eng.* **2023**, *79*, 107904. [CrossRef]

15. Fang, G.H.; Lin, Z.M.; Xie, C.Z.; Han, Q.Z.; Hong, M.Y.; Zhao, X.Y. Optimized Machine Learning Model for Predicting Compressive Strength of Alkali-Activated Concrete Through Multi-Faceted Comparative Analysis. *Materials* **2024**, *17*, 5086. [CrossRef] [PubMed]

16. Pan, B.; Liu, W.; Zhou, P.; Wu, D.O. Predicting the Compressive Strength of Recycled Concrete Using Ensemble Learning Model. *IEEE Access* **2025**, *13*, 2958–2969. [CrossRef]

17. Wang, J.; Deng, J.; Li, S.; Du, W.; Zhang, Z.; Liu, X. Explainable Machine Learning for Multicomponent Concrete: Predictive Modeling and Feature Interaction Insights. *Materials* **2025**, *18*, 4456. [CrossRef] [PubMed]

18. Saleh, M.A.; Kazemi, F.; Abdelgader, H.S.; Isleem, H.F. Optimization-based multitarget stacked machine-learning model for estimating mechanical properties of conventional and fiber-reinforced preplaced aggregate concrete. *Arch. Civ. Mech. Eng.* **2025**, *25*, 185. [CrossRef]

19. Safhi, A. A Comprehensive Self-Consolidating Concrete Dataset for Advanced Construction Practices. *Zenodo* **2024**. [CrossRef]