sustainability

MDPI

Article

Cross-Source Reliability of Chemical-Composition-Based AI Models for Predicting the Ultimate Tensile Strength of Cast Al–Si Alloys: Implications for Sustainable Manufacturing

Tomasz Matula [1,*], Jarosław Piatkowski [2], Silvie Brozova [3] and Beata Oleksiak [4]

[1] Department of Metallurgy and Recycling, Faculty of Materials Engineering and Industrial Digitalization, Silesian University of Technology, 8 Krasińskiego Street, 40-019 Katowice, Poland
[2] Materials Innovation Laboratory, Faculty of Materials Science and Industrial Digitalization, Silesian University of Technology, Krasińskiego 8, 40-019 Katowice, Poland
[3] Department of Metallurgical Technologies, Faculty of Metallurgical Technologies, VSB—Technical University of Ostrava, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, Czech Republic; silvie.brozova@vsb.cz
[4] Department of Production Engineering, Faculty of Materials Science, Silesian University of Technology, Krasińskiego 8, 40-019 Katowice, Poland
* Correspondence: tomasz.matula@polsl.pl

Abstract

This study assesses the cross-source reliability of machine-learning models predicting ultimate tensile strength (UTS) of Al–Si alloys using chemical composition alone. The Zhao training source contained 216 observations but only 156 distinct composition vectors; 32 repeated vectors accounted for 92 rows. For this Zhao dataset and its recorded chemical-input representation, within-group UTS variation yielded a descriptive minimum in-sample RMSE of 23.07 MPa, not a universal accuracy limit. Mean-only, Ridge and Extra Trees models trained on 211 screened Zhao rows were assessed exploratorily on Cai (12 compositions) and Bergmann Kremer (18 processing configurations, three nominal compositions). Extra Trees yielded mean absolute errors of 150.56 and 52.21 MPa, respectively, compared with 121.31 and 87.34 MPa for the Zhao training-mean baseline. Nominal 90% prediction intervals calibrated on Zhao achieved median external coverages of 16.7% and 97.2%, respectively, with a median full width of 178.18 MPa. Reweighting repeated compositions changed individual errors but did not remove source-dependent reliability concerns. An additional 12-condition AlSi17 series illustrated the inability of chemistry-only inputs to represent processing-dependent strength differences. These exploratory findings motivate process-specific validation and cautious withholding of unreliable predictions before industrial application. No environmental benefits were measured.

Keywords: sustainable manufacturing; aluminum–silicon alloys; casting; composition-only modeling; source-dependent errors; machine learning; data provenance; predictive uncertainty

Check for updates

Academic Editors: Ramakrishnan Ramanathan and Usha Ramanathan

Received: 31 October 2026
Revised: 21 November 2026
Accepted: 24 November 2026
Published: 27 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Aluminum–silicon alloys are produced using different casting methods. Depending on the alloy grade, its intended application and the required properties, the resulting castings may subsequently undergo heat treatment, including solution treatment and ageing, or, in selected processes, thermomechanical treatment. The tensile properties of Al–Si alloys depend not only on chemical composition but also on the solidification process, cooling rate, silicon-particle morphology, intermetallic phases, casting defects and the material condition resulting from the applied processing route and any subsequent heat treatment. The

influence of alloying additions on the microstructure and properties of cast Al–Si alloys has been reviewed [1], together with approaches to controlling eutectic-silicon morphology [2]. Studies on gravity and semi-solid casting of Al7SiMg [3] and ablation casting of AlSi7Mg [4] demonstrate relationships between processing conditions, microstructure and properties. The effects of melt thermal treatment on cast Al–Si alloys have also been reviewed [5]. Chemical composition alone does not fully encode these metallurgical influences.

The use of machine learning for the design and characterization of metallic alloys has received increasing attention [6]. Examples of machine-learning applications to aluminum alloys include tensile-strength prediction models and the optimization of material properties as a function of chemical composition [7–9]. More broadly, relationships among chemical composition, microstructure and the mechanical properties of alloys are also important. Zhao et al. [10] reported models of aluminum–silicon alloys developed using chemical and additional engineered information. The downloadable supplementary table provides recorded composition vectors and property responses but does not supply complete processing histories or source-publication identifiers for each row. The present study examines what can reliably be inferred from its chemical inputs alone and does not reproduce or directly evaluate the original broader Zhao approach. Random splitting of literature-derived observation rows may yield overly optimistic results when repeated or closely related chemical compositions occur simultaneously in training and testing. Relevant research on leakage, realistic evaluation and distribution shift in materials informatics highlights the need for appropriate validation units [11–14]. Another difficulty arises when identical recorded inputs correspond to different mechanical properties. A deterministic chemistry-only regressor then necessarily issues the same prediction despite different measured strengths.

Prediction-uncertainty quantification cannot remedy missing material-state information when calibration assumptions fail in the application domain. Residual-calibrated prediction intervals require appropriate exchangeability assumptions, which need not hold across different experimental publications [15–17]. Risk-ranking scores, including chemical distance and disagreement among tree predictions, similarly require direct testing rather than assumptions of transferability [18]. In aluminum foundries, choices of alloy composition, melt preparation, casting technique and heat treatment affect both material properties and resource use. Relevant research includes the selection of alloy compositions according to circular-economy principles, the preparation of melts using secondary raw materials and approaches to reducing the resource intensity of heat-treatment processes [19–21]. However, the present study does not claim that its AI models have reduced energy use, casting rejects or emissions. Instead, it investigates whether composition-only predictions can be considered sufficiently reliable for future materials decision support. The link to sustainable manufacturing therefore follows a testable decision pathway: predictions must first be validated in the target foundry, their influence on acceptance or rejection decisions must then be compared with established practice, and only subsequently can changes in material or energy use be quantified.

The training source includes heterogeneous aluminum-rich compositions, including six records with a reported silicon fraction of zero; its individual casting and processing histories are not fully traceable. Accordingly, this study assesses the reliability of chemistry-only predictions across selected aluminum-alloy sources, with particular implications for Al–Si casting-related applications, rather than claiming coverage of a uniformly defined cast Al–Si alloy population.

The objective of this study is to quantitatively assess the limits of the cross-source reliability of machine-learning models that predict UTS of cast Al–Si alloys solely from chemical composition. The analysis addresses three research questions:

1. To what extent do repeated chemical-composition vectors associated with different UTS values limit the unique mapping of mechanical properties from the recorded inputs?
2. How do predictive accuracy and the direction of bias change when a model trained on one dataset is applied to observations from other publications and processing conditions?
3. Do prediction-uncertainty calibration and selective prediction filters remain useful after transfer between different sources of materials data?

To address these questions, the present study analyzed ambiguity in the training data, performed exploratory comparisons between published sources and examined the transportability of prediction intervals and heuristic risk filters. Previously published AlSi17 results were used as additional empirical confirmation of the limitations of predicting tensile strength from chemical composition alone when processing variables are omitted. The analysis deliberately separates experimental observations, constructed chemical features and missing processing information. The contribution is an integrated quantitative empirical assessment of the limitations of composition-only models in cast Al–Si alloys, rather than a new machine-learning method. The study links input ambiguity, source-dependent predictive errors, uncertainty calibration and the conditions for reliable decision support. Environmental effects and independent prospective industrial validation are outside its measured scope. Thus, the sustainability contribution concerns the reliability conditions for responsible AI-assisted decisions, not a demonstrated improvement in production efficiency or environmental performance.

2. Materials and Methods

2.1. Data Sources and Observation Units

The training source was the Zhao et al. supplementary CSV (multimedia component 2; 216 rows × 49 fields) [10]. Its variables comprise a row index, 17 elemental atomic-fraction columns (including Al), 28 constructed physicochemical descriptors and three mechanical-property columns (tensile strength, yield strength and elongation). The exact 17-element composition vector served as the grouping identifier. Of 216 rows, 156 feature vectors were distinct. An accompanying Word supplement documents a processing-code convention, but the codes and heat-treatment entries are not assigned to the 216 released CSV rows.

The exploratory cross-source datasets comprised: Cai et al. [22], 12 individual source-reported compositions with experimentally measured ultimate tensile strength (nine gravity-cast + T6; three gravity-cast + hot-extruded). For Cai, model chemistry inputs were taken from the compositions tabulated for the alloy-design iterations; these twelve input vectors should not be presented as twelve independently documented, complete chemical assays. The original study reports separate spectroscopic checks for selected alloys. The second exploratory source was Bergmann Kremer et al. [23], 18 reported as-cast process configurations across AlSi9Cu3, AlSi10 and AlSi12. The latter 18 observations are represented here by three source-tabulated nominal/OES chemical vectors (one per alloy), rather than by measured composition for each individual heat or treatment. Thus, the two sources have markedly different units of chemical independence. No row-level provenance identifiers were available for the Zhao records; exact chemical non-overlap does not certify non-overlapping laboratories or experiments. A third, separately classified series was Hillebrandt-Szymańska and Piątkowski [24], with 12 reported means from an AlSi17 experiment (six melt-overheating temperatures × two mould types). This author-affiliated source contributes one reconstructed chemical input vector, not 12 independent compositions. It was not used for model training, tuning or calibration; the observed labels had already been examined during project development.

Distinct chemical-composition vectors were identified by exact numerical comparison of all 17 recorded elemental atomic fractions using NumPy unique along the observa-

tion axis. No chemical-similarity threshold, rounding tolerance or clustering was introduced. The 216 records contained 156 distinct recorded vectors; 32 of these recurred in 92 observations. After excluding five records flagged by the cross-property quality screen, 211 observations and 151 distinct vectors remained, whereas the 32 repeated groups still comprised 92 observations. These groups are identical recorded inputs, not experimentally verified independent melts or casting batches. Small differences in reported precision may yield separate vectors. In internal data partitioning, all records with the same exact input vector were assigned to the same partition; this avoids direct duplicate-input overlap but cannot establish independent experimental provenance.

The silicon-content ranges and observation units are summarized in Table 1. Silicon content in the 216-record Zhao supplementary data [10] ranged from a reported atomic fraction of 0 to approximately 0.20649 (approximately 0.00–21.31 wt.% after composition-based conversion). Six records had Si recorded as zero and remained among the 211 screened records. Thus, the source cannot be characterized as a homogeneous set of exclusively cast Al–Si alloys. Cai [22] covered 4.00–12.30 wt.% Si; the three nominal Bergmann Kremer [23] alloys contained 8.52, 9.62 and 11.84 wt.% Si; and the AlSi17 source [24] contained 16.63 wt.% Si. These published ranges are descriptions of the source inputs, not validated industrial prediction domains.

Table 1. Dataset scope, silicon-content ranges and units of independence. Nominal compositions, processing configurations and experimentally independent melts are distinct observational units.

Source	Number of Observations/ Input Groups	Role and Material Condition	Si (wt.%)
Zhao [10]	216 records/156 unique compositions	Training; heterogeneous processing histories unavailable at record level	0.00–21.31
Cai [22]	12 source-tabulated compositions	Exploratory cross-source assessment; nine gravity-cast + T6 variants, three gravity-cast + extruded variants	4.00–12.30
Bergmann Kremer [23]	18 processing conditions/ 3 nominal compositions	Exploratory cross-source assessment; as-cast (F); sand mold versus metal mold	8.52–11.84
Hillebrandt-Szymańska and Piątkowski [24]	12 process-condition means/1 reconstructed composition	Exploratory AlSi17 series from a coauthor publication; six temperatures × sand/copper mold; processing variables absent from model inputs	16.63

The 16 source-reconstructed external chemical input vectors, showing both reported values and the computational zeros used for unreported elements, are supplied in Appendix Table A4. The full 216-record original Zhao matrix is available in the source supplementary file [10] and is not republished here.

The scope of the datasets and their units of observation are summarized in Table 1.

The mapping of source-reported elemental compositions to the 17 model inputs is shown in Table 2.

Element mass percentages reported in the external publications were converted to atomic fractions according to $x_i = (w_i/M_i)/\Sigma_j(w_j/M_j)$, where w is the mass fraction used for reconstruction and M the atomic mass. The model inputs were ordered exactly as in the Zhao et al. supplementary dataset [10]: Al, B, Cr, Cu, Er, Fe, Mg, Mn, Ni, Sc, Si, Sn, Sr, Ti, V, Zn and Zr. Table 2 lists the source-to-feature mapping for all four evaluated datasets. When aluminum was reported as the balance, its mass fraction was calculated by difference from 100%. Any element among the model inputs that was not quantified in a source was assigned zero for computational reconstruction. This convention does not

establish analytical absence in the actual alloy. Neither the aluminum balance nor the zero assignment represents an independent chemical measurement; the adopted assumptions can increase uncertainty in cross-source predictions. In reconstructing the Cai data [22], the mass percentages of Al, Si, Mg, Cu, Ni, Zn, Fe, Mn and Cr were transcribed from the separate columns of the published Table 1. Both Zn and Cr were reported as zero for all twelve selected alloy compositions. Table 1 contains no aggregate "other elements" column, and no source-reported residual was reassigned to Zn. Additional Zhao model elements not reported in Cai Table 1 were encoded as zero solely as a computational convention; this does not establish their analytical absence. The input vectors were based on the tabulated compositions rather than twelve separately documented full compositional assays. The unusual Mg content reported for Cai sample D3 [22] was retained without alteration in the main analysis; its influence was examined separately in a sensitivity check. No published source value was replaced with a corrected or inferred measurement.

Table 2. Source-to-feature mapping of the 17 elemental inputs. "Reported" indicates a source-tabulated value (Zhao: atomic fraction; external publications: mass fraction before conversion); "Balance" denotes aluminum calculated by difference; "0" denotes an unreported model element assigned zero solely for computational reconstruction. Cai Zn and Cr are source-reported zeros. A zero assignment is not evidence of analytical absence.

Element	Zhao	Cai	Bergmann Kremer	AlSi17
Al	Reported	Reported	Balance	Balance
B	Reported	0	Reported	0
Cr	Reported	Reported	0	0
Cu	Reported	Reported	Reported	Reported
Er	Reported	0	0	0
Fe	Reported	Reported	Reported	Reported
Mg	Reported	Reported	Reported	Reported
Mn	Reported	Reported	Reported	Reported
Ni	Reported	Reported	0	Reported
Sc	Reported	0	0	0
Si	Reported	Reported	Reported	Reported
Sn	Reported	0	0	0
Sr	Reported	0	Reported	0
Ti	Reported	0	Reported	0
V	Reported	0	0	0
Zn	Reported	Reported	Reported	0
Zr	Reported	0	0	0

For the AlSi17 series of Hillebrandt-Szymańska and Piątkowski [24], the input vector was reconstructed using the mass percentages of Si, Cu, Fe, Mn, Mg and Ni reported in that publication. The aluminum mass percentage was determined as the difference between 100% and the sum of the reported mass percentages of these elements. The remaining unreported model elements were set to zero as a computational convention. The 0.05 wt.% phosphorus addition reported in the publication was not included because phosphorus was not among the Zhao model features. The same reconstructed input vector was assigned to all 12 temperature–mold combinations. The reported strength values are process-condition means, not raw measurements from independent tensile specimens.

2.2. Data Quality Flags, Baseline Models and Grouping

Five records in the Zhao et al. supplementary dataset [10] (zero-based row indices 11–15) exhibited reported yield strengths (YS) greater than their corresponding ultimate tensile strengths (UTS). This cross-property inconsistency was treated as an indication of a potential data-quality issue, not as evidence that the reported UTS values were incorrect.

The 211 remaining records formed the primary reported training subset. To assess the sensitivity of this precautionary decision, the same fixed Extra Trees model was additionally fitted to all 216 records, without changing the external Cai or Bergmann Kremer evaluation targets or the model hyperparameters. Original source measurements were not corrected or replaced. The composition-only models were: (a) the mean training-label predictor; (b) Ridge regression with StandardScaler and $\alpha = 10$; and (c) ExtraTreesRegressor with n_estimators = 200, min_samples_leaf = 2, max_features = 0.85 and random_state = 88. All 17 elemental atomic fractions were input; the derived descriptors were not used in the principal cross-source tests. These were fixed configurations rather than the result of a documented grid search or randomized hyperparameter search. Their selection was exploratory, and the reported settings are not claimed to be globally optimal. No external UTS measurements were used to fit these models or to modify their predictions. This protocol differs from the ensemble and feature set used by Zhao et al. [10].

Exact recorded composition vectors served as group identifiers in internal data partitions, preventing identical chemical inputs from appearing in both training and evaluation subsets. An initial comparison of the fixed mean-only, Ridge and Extra Trees configurations used twelve GroupShuffleSplit partitions with a 20% held-out share of distinct chemistry groups (random seeds 881–892). This comparison was not a hyperparameter-optimization procedure. Separately, the subsequent uncertainty-calibration and internal risk-ranking analyses used twelve chemistry-grouped partitions with a 25% held-out share (random seeds 9001–9012). These proportions refer to recorded input groups, not to independent melts, experiments, production batches or publications. The cross-source models were trained on all 211 screened Zhao observations and were not fitted to Cai, Bergmann Kremer or AlSi17 strengths. External predictions were assessed using MAE, RMSE, R^2 and mean signed error (prediction minus observation), with the reported observation or processing condition as the evaluation row. As these publications and their outcomes were available during study development, the external comparisons are exploratory cross-source stress tests rather than blinded prospective validation.

To assess sensitivity to unequal representation of repeated recorded compositions in Zhao et al. [10], an additional post-review analysis compared three training strategies on the same 211 quality-screened Zhao observations: (A) observation-level training with equal row weights; (B) one observation per distinct chemical-composition group, with its target defined as the arithmetic mean UTS of that group (151 groups); and (C) observation-level training with sample weights inversely proportional to composition-group size, normalized to a mean weight of one. For strategy C, sample weighting was applied to both StandardScaler fitting and the Ridge loss, while Extra Trees used weighted fitting. The mean-only comparator used the corresponding training-label mean or weighted mean. The three strategies retained the previously reported fixed Ridge and Extra Trees hyperparameters and were assessed using the original twelve composition-grouped 80:20 Zhao splits (seeds 881–892), as well as the unchanged Cai and Bergmann Kremer external observations. External target measurements were not used to tune models or select training weights. Strategy B changes the training target representation but does not create new independent chemical measurements.

For the within-source Zhao reference, held-out predictions were summarized from the twelve existing composition-grouped splits (seeds 9001–9012; GroupShuffleSplit test_size = 0.25), using the same Extra Trees specification. For each split, MAE, RMSE and R^2 were calculated across held-out observation rows. A separate group-balanced MAE was obtained by first averaging the absolute error within each distinct held-out composition group and then giving each group equal weight. Medians across overlapping splits are descriptive summaries rather than independent replication estimates. Per-split aggregates

are provided in Table A1; the third-party cross-source evaluations instead used models fitted on all 211 screened Zhao records.

2.3. Empirical Ambiguity Between Chemical Composition and Tensile Strength

For each exact 17-dimensional composition vector g, let y_i be the measured tensile strength of observation i in that group and $\bar{y}_g$ the corresponding group average. For each unique chemical-composition vector, a deterministic model using only these inputs must assign the same prediction to every corresponding observation. Because the arithmetic group mean minimizes the sum of squared deviations within that group, the empirical minimum root-mean-square fitting error for the analyzed dataset can be obtained from the deviations of individual UTS values from their respective group means, as expressed in Equation (1):

$$\mathrm{RMSE}_{\min} = \sqrt{\frac{1}{N}\sum_{g}\sum_{i\in g}\left(y_i - \bar{y}_g\right)^2} \quad (1)$$

Here, N is the number of rows and g indexes groups of identical recorded chemical inputs. For all 216 Zhao observations (156 different vectors), this descriptive minimum was 23.07 MPa. After screening, the corresponding value for 211 observations (151 vectors) was 23.34 MPa. All five flagged inputs were singletons, so their removal left the within-group residual sum of squares unchanged but reduced the normalizing sample size. These bounds quantify feature–response ambiguity in the observed table; they do not represent an irreducible physical noise level, a limit on generalization to new alloys or a guaranteed performance of a particular trained algorithm. Ambiguity may reflect unrecorded processing or structural conditions, measurement variability, reporting precision or other factors. The available records cannot distinguish these possibilities or determine what fraction of the observed variability could be explained by additional process measurements.

2.4. Repeated Uncertainty Calibration and Risk Ranking

To assess the transfer of prediction uncertainty between sources, twelve splits of the Zhao et al. [10] train/calibration partitions were generated using GroupShuffleSplit (test_size = 0.25; random seeds 9001–9012). Every observation with an identical recorded 17-element chemistry vector was retained within the same partition. Only the 211 records passing the predefined quality screen were used. A separate Extra Trees model was trained for each split using only its assigned Zhao training records. The remaining calibration groups did not participate in fitting. The same external observations were subsequently evaluated under all twelve dependent splits. Thus, the reported median scores summarize repeated internal calibration, not twelve independent external datasets. Prediction-interval calibration. Nominal 90% intervals were constructed using a chemistry-grouped adaptation of the split-conformal procedure [15,16]. For each distinct composition group in the calibration partition, the maximum absolute prediction residual among that group's observations was calculated. The half-width of the symmetric interval was selected as the order statistic of the sorted group-maximum residuals at rank $\lceil (G_{cal} + 1) \times 0.90 \rceil$, where G_{cal} is the number of distinct calibration composition groups. If this rank exceeded the number of available groups, the formal interval width was set to infinity.

This procedure treats chemical groups, not independent individual rows, as calibration units. Any interpretation of a nominal coverage guarantee requires appropriate exchangeability assumptions at the group level. These assumptions were not established for observations from different publications and processing conditions. Therefore, the nominal 90% level does not guarantee equivalent coverage for Cai or Kremer. Coverage was computed empirically for each split and summarized by the median of the twelve repetitions; repeated use of the same external observations does not constitute twelve in-

dependent validations. Under the stated exchangeability requirement, group-maximum calibration targets prediction performance for a new exchangeable chemistry group; it does not establish 90% coverage for an individual observation from an arbitrary new publication or process. Selective acceptance of predictions. Independently of interval coverage, two heuristic risk scores were evaluated: (a) chemical distance to the nearest observed training composition, and (b) standard deviation of predictions from the individual trees of the Extra Trees ensemble. Distance was calculated in the standardized space of the 16 elemental atomic fractions excluding Al, whose amount is constrained by the composition sum. Scaling parameters and the nearest-neighbor structure were learned from training observations only. Across-tree spread was treated as a measure of disagreement within the ensemble rather than a separately calibrated probability of prediction error.

For each score, the 80th and 95th percentiles across the distinct composition groups in the calibration partition defined acceptance thresholds. A prediction was accepted when the score did not exceed the chosen threshold. The distance filter, the across-tree spread filter and their conjunction, which required both criteria to be satisfied, were examined. Acceptance thresholds were determined without access to the UTS observations in the external sources. Those measurements were used only subsequently to evaluate predictive errors and filter performance. Calibration percentiles of 80 and 95% do not define nominal prediction-interval coverage and do not guarantee the same percentages of accepted external records. Accepted processing configurations were distinguished from distinct accepted compositions. When all predictions were rejected, conditional MAE remained undefined. With respect to internal risk ranking, in a separate study restricted to Zhao, a score based on local variation in UTS was assessed. Twelve chemistry-grouped splits with the same random seeds and approximately 25% held-out composition groups were used. Each Extra Trees model was fitted exclusively on its corresponding training portion.

For each unique training composition, within-group mean and variance of UTS were computed. Chemical neighborhoods were constructed using seven elements: Si, Cu, Mg, Fe, Mn, Ni and Ti. After standardization based on unique training compositions, the $k = 4$, 8 or 12 nearest distinct training groups were selected. Neighbor weights were inversely proportional to $0.25 + d_j$, where d_j is the standardized Euclidean distance, and normalized to sum to one. The local total-variation score $\mathrm{LTV_k}$ combined the weighted within-training-group variance of UTS and the variation in group mean UTS within the selected chemical neighborhood. The following definition was used: Local total variation was defined as $\mathrm{LTV}_k(x) = \left\{\sum_j w_j \left[s_j^2 + \left(\mu_j - \bar{\mu}_w\right)^2\right]\right\}^{1/2}$, where μ_j and ${s_j}^2$ are the mean and variance of UTS for the jth neighboring training group, and $\bar{u}_w = \Sigma_j\, w_j\, \mu_j$. Thus, the score combines within-composition variation with differences between the mean responses of neighboring compositions. The reference score was the standard deviation of individual tree predictions for the same input. For each unique held-out composition group, its average absolute prediction error was calculated from the corresponding test rows. Every distinct group received equal weight regardless of its number of records. Prediction-risk ranking was assessed using the area under the risk–coverage curve (AURC), a measure employed in selective prediction [18]. A discrete, group-balanced definition was adopted that assigns equal weight to each distinct chemical-composition vector, irrespective of the number of observations in that group. After ordering the groups by increasing risk score, AURC was calculated as the mean of the cumulative MAE values over successive group-acceptance fractions:

The seven elements Si, Cu, Mg, Fe, Mn, Ni and Ti were used only to define chemical neighborhoods in the exploratory local total-variation (LTV) risk score. The principal Ridge and Extra Trees UTS predictors retained all 17 elemental atomic-fraction features derived from Zhao et al. [10]. The shorter list must not be interpreted as a different feature set

for the main predictive models. Element lists in external source tables instead reflect the varying composition information supplied in those publications.

$$\mathrm{AURC} = \frac{1}{G}\sum_{j=1}^{G}\left(\frac{1}{j}\sum_{l=1}^{j} e_{(l)}\right) \quad (2)$$

Here, G denotes the number of distinct held-out composition groups and $e_{(l)}$ their mean absolute error at ordered position l. Discrete AURC is expressed in MPa; smaller values indicate more favorable aggregate risk ranking. MAE_{80} was additionally defined as the mean error of the first $\lceil 0.8G \rceil$ ranked held-out groups. Accepting 80% of groups is not equivalent to using the 80th-percentile risk threshold in the cross-source filtering experiment. This group-balanced AURC should not be compared directly with the row-weighted empirical minimum RMSE in Section 2.3: they use different loss functions, aggregation units and evaluation settings. Neighborhood sizes and risk criteria were examined exploratorily in overlapping splits of the same Zhao dataset. They are not independent experimental replications and cannot establish superiority of the proposed score. The local risk ranking was not validated on a separate external source. Calibration-transfer and selective-acceptance analyses were performed for the Cai and Bergmann Kremer data. The additional AlSi17 case was assessed separately as an illustration of processing effects unavailable to composition-only predictions; it was not used to choose thresholds or establish calibration coverage.

2.5. Sustainable Manufacturing Decision Context and Scope of Assessment

A potential use of machine-learning predictions of tensile strength is preliminary screening of Al–Si alloy compositions before full experimental qualification. Such use would require a specified target casting process, material condition and mechanical acceptance criteria. A chemistry-only model cannot distinguish processing configurations that share an identical composition vector but differ in overheating temperature, cooling method, mold type or heat treatment. For decision-support purposes, at least two kinds of predictive error should be distinguished. Underestimating UTS may lead to the rejection of an otherwise acceptable alloy or unnecessary verification, while overestimating UTS could result in premature acceptance of unsuitable material. Their consequences depend on actual acceptance limits, verification procedures and the processing system; MAE alone does not determine the rate or cost of incorrect decisions.

This study assessed ambiguity between chemistry and UTS, error magnitude and signed bias across sources, empirical prediction-interval coverage and exploratory filtering of predicted responses. It did not implement a decision rule in a production foundry and did not measure casting rejects, energy consumption, remelting mass, resource savings or greenhouse-gas emissions. A future sustainability-oriented assessment of these models should specify a casting process, minimum mechanical requirements, the relevant unit of production, decision thresholds and the competing reference procedure. Prediction and abstention rules should then be tested on genuinely new casting series, alongside resource-use and environmental indicators. Accordingly, the present results identify reliability and input-data requirements that would need to be satisfied before using composition-based predictions as evidence for sustainable foundry decision support. They do not establish measured improvements in environmental performance.

2.6. Computational Reproducibility

All analyses were performed using Python scripts while keeping the Zhao training data separate from the published Cai and Bergmann Kremer comparison sources. The environment used for the computational reproducibility check comprised Python 3.13.5, NumPy 2.3.5, scikit-learn 1.8.0 and SciPy 1.17.0. Model settings, chemistry reconstruction conventions, data splitting and assessment criteria are documented in the Methods and

the accompanying code. To assess computational reproducibility, cleaned scripts were run again in a separate working directory using the original Zhao CSV and locally reconstructed Cai and Bergmann Kremer inputs. The additional AlSi17 study based on Hillebrandt-Szymańska and Piątkowski [24] was checked separately using the same fixed model and reconstructed chemistry input. The 12 source means were not used for training, and this auxiliary check was not included among the ten tables of the principal reproducibility audit. They do not constitute independent experimental validation, nor do they independently establish the accuracy of transcription or the right to republish third-party observations. The review package contains author-developed Python scripts, empty source-data templates, execution instructions, pinned dependency versions and derived aggregate outputs. Original observation-level third-party tables are excluded. Public release is subject to author approval and source-data reuse requirements.

The reproducibility audit covered ten numerical output tables generated during the exploratory cross-source assessment, prediction-interval calibration, and risk-assessment analyses. These included individual predictions and aggregated performance metrics. The reproduced results were compared with the archived numerical outputs using absolute and relative tolerances of 1×10^{-9}. No discrepancies exceeding these thresholds were identified, and the maximum absolute numerical difference was approximately 1.14×10^{-13}. The audit verified computational consistency under the specified implementation and data-reconstruction procedures.

2.7. Hypothetical Zero-Assignment Sensitivity and Process-Aware Comparison

To assess input-reconstruction sensitivity, the original 211-record Zhao-trained Ridge and Extra Trees models and their fixed hyperparameters were retained without refitting. Each external composition was reconstructed under hypothetical individual additions of 0.01 or 0.10 wt.% of an element not quantified in the relevant source, and joint additions of 0.10 or 0.50 wt.% distributed equally over all unreported model elements. The aluminum balance was reduced correspondingly to preserve 100 wt.%, and all resulting mass fractions were converted back to the same 17 atomic-fraction features. Source-reported zeros, including Zn and Cr in Cai, were not treated as missing values. For each scenario, prediction differences relative to the original zero-encoded inputs were calculated; for complete joint scenarios, condition-level MAE was additionally compared. These are deterministic what-if perturbations, not analytical detection limits or estimates of actual unmeasured element concentrations. Phosphorus reported for AlSi17 was not varied because P is absent from the Zhao feature set.

For the additional processing analysis, original Zhao-trained predictions for Cai were separated into nine gravity-cast/T6 and three gravity-cast/hot-extruded conditions; Bergmann Kremer predictions were separated by mold type and melt-preparation category. Because the Zhao training rows lack linked process metadata, a paired chemistry-only versus chemistry-plus-process model could not be fitted on Zhao. A separate exploratory leave-one-composition-out (LOCO) comparison was therefore performed on the three Bergmann Kremer nominal alloys (18 conditions): in each of three folds, both feature representations were trained on twelve conditions belonging to two compositions and tested on all six conditions of the third. The additional categorical process features encoded three melt-preparation categories and two mold categories within the training fold. Fixed Ridge (StandardScaler; alpha = 10) and Extra Trees (200 trees, minimum leaf size 2, maximum features 0.85, random state 88) settings were applied in both representations. The LOCO exercise is a within-Kremer, very-small-sample experiment, not validation of a Zhao-trained process-aware model or an independent foundry-scale generalization test.

3. Results

3.1. Research Question 1: Recorded Input–Response Ambiguity

Analysis of the supplementary dataset published by Zhao et al. [10] showed that the 216 observations represented only 156 distinct chemical-composition vectors. Thirty-two vectors occurred more than once, accounting for 92 records in total. Thus, the number of distinct model inputs was substantially smaller than the number of rows. Different reported UTS values were observed within repeated-composition groups. Because a deterministic regression model receives the same input for every record in such a group, it must assign the same prediction to all of them. Consequently, some residual variation remains even for a hypothetical model that exactly fits the mean response of every observed group. As defined in Section 2.3, the sum of squared deviations of UTS from its corresponding group means was 114,969.03 MPa^2. Dividing by the full sample size of 216 and taking the square root yielded the descriptive minimum RMSE of 23.07 MPa. This is the best possible in-sample fit for a deterministic predictor that assigns one constant to each exact input group. After excluding five records flagged during quality control, the main dataset contained 211 observations and 151 distinct composition vectors. All five excluded records were singleton groups; hence, the within-group sum of squares was unchanged, whereas the denominator decreased. The corresponding descriptive minimum RMSE was 23.34 MPa.

These dataset-specific values describe input–response ambiguity in the Zhao observations under the recorded 17-element chemical-input representation. They must not be interpreted as prospective test errors, irreducible physical noise, or universal accuracy limits for models that include additional information about the alloy and its processing history. Variation in UTS associated with identically recorded chemistry may be caused by differences in processing routes, microstructure, casting defects, heat-treatment state, test conditions, measurement variability or the precision of the reported inputs. The available data do not allow the individual contributions of these factors to be estimated. Within-source performance on the twelve Zhao composition-grouped holdouts provides a reference for the externally evaluated predictions. Each holdout contained 38 distinct chemical-input groups (44–67 observation rows). Extra Trees yielded a median row-weighted MAE of 29.62 MPa, RMSE of 42.35 MPa and R^2 of 0.694; the median composition-group-balanced MAE was 24.69 MPa (Table A1). Individual-split row-weighted MAE ranged from 21.88 to 52.37 MPa and R^2 from −0.116 to 0.851, with negative R^2 in two splits. The repeated holdouts overlap and the external-source predictions were obtained from models trained on all 211 screened Zhao records, rather than on the restricted within-source training subsets. Consequently, these metrics are a descriptive within-source reference and do not isolate the causal effect of changing data sources. In the separate earlier 80:20 grouped comparison (twelve splits, seeds 881–892), the median held-out MAEs were 66.08 MPa for the mean-only baseline, 43.71 MPa for Ridge and 26.98 MPa for Extra Trees. These are descriptive internal comparisons of fixed configurations, not hyperparameter-selection results or independent prospective validation.

3.2. Research Question 2: Cross-Source Predictive Reliability and Its Sensitivity

The Extra Trees model trained on the 211 screened Zhao records and supplied exclusively with the 17-element chemistry vector exhibited substantial differences in predictive performance between the two published external datasets (Table 3 and Figures 1 and 2). The fixed model, Ridge regression and the training-mean baseline were evaluated without refitting on the external measurements. For Cai, which comprised 12 source-reported compositions, Extra Trees yielded MAE = 150.56 MPa, RMSE = 161.66 MPa and $R^2 = -1.84$. The mean signed error, defined as prediction minus observation, was −120.77 MPa, indicating pronounced average underprediction. The mean-only predictor gave a lower MAE

of 121.31 MPa, whereas Ridge yielded 152.37 MPa. The Cai observations included distinct chemical compositions and different material processing states, including gravity casting followed by T6 treatment and configurations involving extrusion. The analysis did not isolate the contribution of any individual process or alloying element to the observed differences.

Table 3. Exploratory results for Extra Trees, the mean-only baseline and Ridge on two previously examined publications, after training on 211 Zhao records.

Metric/Baseline Model	Cai: 12 Compositions	Bergmann Kremer: 18 Conditions (3 Compositions)
Extra Trees MAE (MPa)	150.56	52.21
Extra Trees RMSE (MPa)	161.66	59.17
Extra Trees R^2	−1.84	−6.90
Mean signed error (MPa)	−120.77	+50.92
Mean-only predictor MAE (MPa)	121.31	87.34
Ridge MAE (MPa)	152.37	43.01

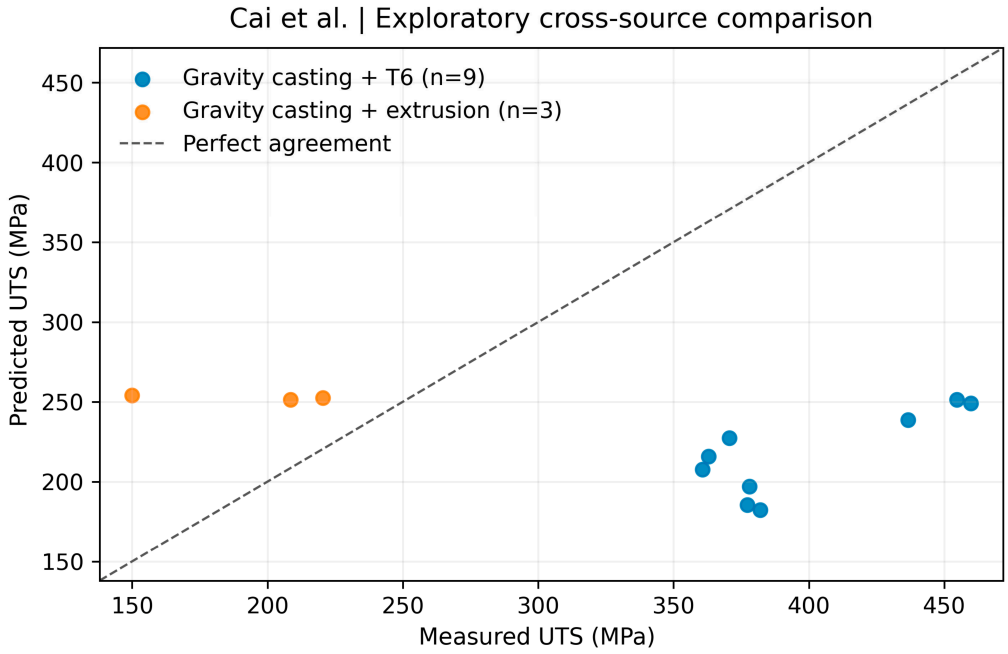

Figure 1. Extra Trees predictions (model trained exclusively on Zhao) versus measured UTS for Cai et al. [22]. The scatter plot is an exploratory cross-source stress test including different processing states.

For Bergmann Kremer, the 18 processing configurations corresponded to just three nominal compositions. Extra Trees yielded MAE = 52.21 MPa, RMSE = 59.17 MPa and R^2 = −6.90. The mean signed error was +50.92 MPa, indicating systematic overprediction on average. For comparison, the training-mean predictor yielded MAE = 87.34 MPa and Ridge MAE = 43.01 MPa. Unlike the results for Cai, Extra Trees achieved an MAE lower than the mean-only baseline by 35.13 MPa (40.2%) for Kremer. However, Ridge performed better than Extra Trees on this source. The relative ranking of the evaluated regression methods therefore differed between the two external publications and cannot be interpreted as a general advantage of any one model class. For Bergmann Kremer, the 18 configuration-level errors arise from only three nominal chemical inputs and must not be interpreted as 18 independent demonstrations of model ranking.

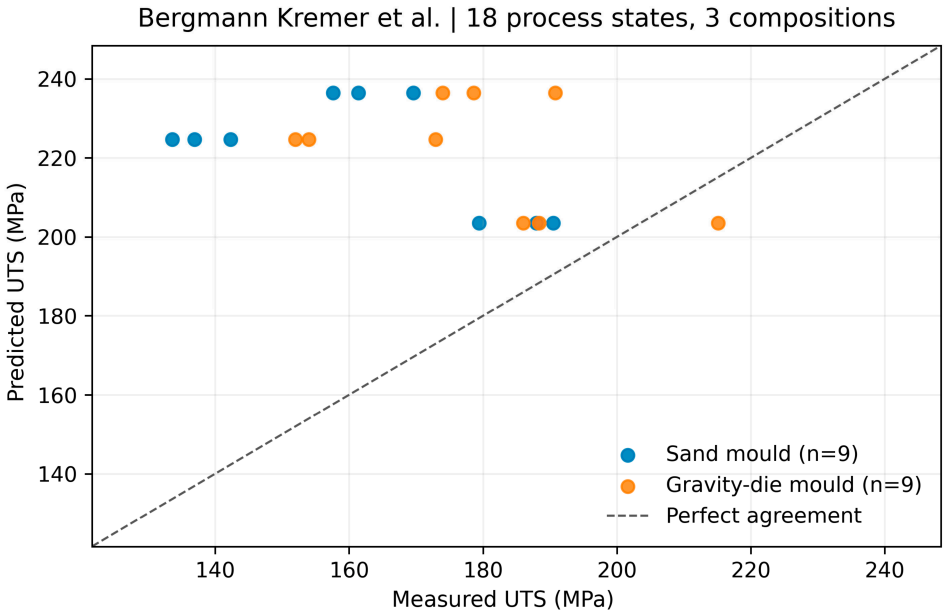

Figure 2. Extra Trees predictions versus measured UTS for Bergmann Kremer et al. [23]. Six processing configurations for each of three nominal chemical vectors share the same model input; the 18 points are not chemically independent.

The negative R^2 values should be interpreted relative to the range and variance of UTS within each external source: this statistic compares the sum of squared errors with the source-specific variation around its own mean. It is not directly interchangeable with the MAE comparison to a baseline trained on Zhao, nor does a negative R^2 alone identify the physical cause of poor transfer. A supplementary sensitivity analysis showed that training the same Extra Trees algorithm on all 216 Zhao records reduced external MAE to 150.19 MPa for Cai and 39.60 MPa for Bergmann Kremer. The difference from the screened 211-record variant was small for Cai but amounted to 12.61 MPa for Kremer. These results show that removing five quality-flagged observations may affect the two external sources differently. These comparisons should be regarded as exploratory assessments of the reliability of models trained using chemical composition alone. The Cai and Bergmann Kremer sources were examined during project development and therefore are not prospective blinded validation datasets. Moreover, Kremer's 18 processing configurations represent just three nominal chemical inputs. The reported errors do not directly assess the original process-aware Zhao model and do not demonstrate causal effects of specific processing parameters on predictive accuracy.

3.2.1. Processing Differences Not Captured by Chemistry-Only Inputs

The data published by Bergmann Kremer et al. provide a particularly clear illustration of a limitation inherent to regression models that use only chemical composition. Processing configurations of the same nominal alloy can exhibit different measured UTS values while sharing identical input features in the reconstructed dataset. Nine reported pairs were compared for the same nominal alloy family and melt-treatment category, contrasting sand-mold and permanent-mold casting. In each pair, the metal-mold condition had higher UTS; the mean paired difference (metal mold minus sand mold) was +16.94 MPa, with a range of +0.4 to +30.6 MPa. Mean UTS values were 162.14 and 179.09 MPa for the two respective processing categories. Because these nine contrasts involve processing variations across just three nominal alloy chemistries, the reported mean difference is descriptive for those configurations, not an estimate based on nine independently sampled alloys. Under the adopted input reconstruction, configurations representing one nominal alloy were assigned

the same 17-element chemical-composition vector. Thus, the 18 processing configurations represent only three nominal chemical inputs, not 18 independent chemical-composition samples. A deterministic composition-only predictor necessarily produces identical UTS estimates for configurations associated with the same vector, notwithstanding their different observed strengths.

This discrepancy does not result from a programming error in the regression algorithm but from information missing from its inputs. A model based on chemical composition alone cannot distinguish differences in mold type or other processing parameters that have not been supplied as features. Nevertheless, the comparison does not isolate the causal influence of mold material from other potentially differing factors. For completeness, a hypothetical minimum training error was calculated for a deterministic predictor assigning a single mean strength to each of the three nominal chemical inputs, using all 18 observed responses. The resulting in-sample RMSE was 11.82 MPa. This hindsight diagnostic uses the external target values to fit the predictions; it is not a legitimate estimate of prospective predictive performance. This analysis indicates that differing technological conditions within the same nominal alloy family can correspond to differing mechanical responses while remaining invisible to a chemistry-only regressor. It does not establish the independent causal effect of any single process parameter.

3.2.2. Additional AlSi17 Processing-Series Illustration

A further exploratory assessment used the results published by Hillebrandt-Szymańska and Piątkowski [24] for an AlSi17 alloy, examining the effects of melt-overheating temperature and mold type. The publication reports 12 condition-level mean UTS values corresponding to six overheating temperatures (760–960 °C) and two mold types (sand and copper). This previously published series was not newly cast or tested for the present study. The reported alloy composition was Si 16.63 wt.%, Cu 0.07 wt.%, Fe 0.34 wt.%, Mn 0.03 wt.%, Mg 0.04 wt.% and Ni 0.02 wt.%, with Al as the balance. The chemistry was converted to a single model input vector. The 0.05 wt.% phosphorus addition was omitted because P was not among the 17 Zhao features, while unreported elements were assigned zero as an encoding assumption. To provide qualitative microstructural context, Figure 3 reproduces three published optical micrographs from the same source [24]. The reference material without overheating (a), shown separately from the six processing temperatures analyzed above, contains relatively large and irregular primary silicon crystals. With overheating to 840 °C (b) and 920 °C (c), the published images illustrate changes in crystal size, shape and distribution, with more compact morphologies evident at the higher temperature. These pre-existing micrographs were neither generated nor quantitatively reanalyzed in the present study. They serve only as a physical illustration of processing-related structural variability; microstructure and processing temperature are not features of the chemistry-only prediction model.

The Extra Trees model trained on 211 Zhao observations, without using AlSi17 responses during fitting, returned the identical prediction UTS = 177.13 MPa for all 12 conditions. In contrast, the published condition means ranged from 198 to 294 MPa, giving MAE = 72.79 MPa and RMSE = 79.58 MPa. Ridge yielded MAE = 52.44 MPa and the Zhao training-mean predictor MAE = 28.91 MPa. Measured UTS varied markedly with processing conditions even though the model could not distinguish the corresponding inputs. Consequently, variations of overheating temperature and mold type were not represented in the prediction and should not be interpreted as quantities learned by the chemistry-only model. In the six paired comparisons at the same overheating temperature, copper-mold UTS exceeded sand-mold UTS in every case. The respective differences were 11, 21, 13, 19, 15 and 14 MPa, with an arithmetic mean of +15.50 MPa. Because both members of each

pair were assigned the same reconstructed chemical input, their model predictions were also identical. A hypothetical fit using one constant equal to the mean of the 12 known published condition responses yielded RMSE = 32.16 MPa. This is a retrospective in-sample description of the variation of reported condition means under a single reconstructed input, not a measured irreducible error or an independent external prediction bound. These results illustrate the limits of a material representation based only on chemistry: a model cannot directly recover differences in processing conditions that are absent from its features. The source was published by a coauthor of the current article and was available during project development. It must therefore be treated as a complementary illustration rather than as independent, blinded validation. Nor do the data establish that mold type alone caused the observed strength differences.

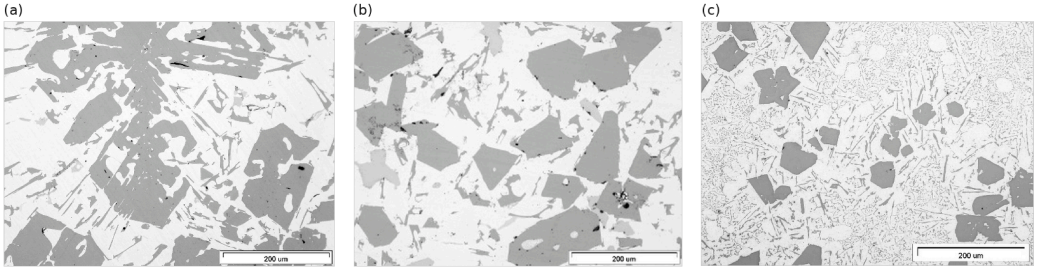

Figure 3. Published optical microstructures of AlSi17: (**a**) reference sample without melt overheating; (**b**) melt overheated to 840 °C; and (**c**) melt overheated to 920 °C. The non-overheated reference is separate from the six overheating temperatures used for the 12-condition UTS comparison. Panels extracted from Figure 4a–c of Hillebrandt-Szymańska and Piątkowski (2025) [24] and rearranged horizontally without modification of the microscopic content; original scale bars retained. © The Authors 2025, reproduced under the Creative Commons Attribution 4.0 International license (https://creativecommons.org/licenses/by/4.0/ (accessed on 23 November 2026)). These are previously published images, not newly collected observations or independent model validation.

The AlSi17 condition-level results and associated chemistry-only model comparisons are summarized in Table 4.

Table 4. Exploratory study of processing variability in the AlSi17 alloy from a coauthor publication [24], using unchanged composition-only models trained on 211 Zhao observations. Twelve published condition means share one reconstructed chemical input; they are not independent chemical compositions.

Parameter	Value	Interpretation
Condition means/reconstructed inputs	12/1	Six temperatures × two mold types
Measured UTS range	198–294 MPa	Means for processing conditions
Extra Trees prediction (every condition)	177.13 MPa	One identical input vector
Extra Trees: MAE/RMSE	72.79/79.58 MPa	Without model refitting
Mean-only/Ridge MAE	28.91/52.44 MPa	Same 211-record Zhao training subset
Group-mean fit RMSE	32.16 MPa	Descriptive; not an external test error
Mean copper minus sand difference	+15.50 MPa	Six pairs at equal temperatures

3.2.3. Process-Stratified Errors and Exploratory Process-Aware Validation

For the nine Cai gravity-cast/T6 compositions, the unchanged Zhao-trained Extra Trees model gave MAE = 180.89 MPa and mean signed error = −180.89 MPa; the three gravity-cast/hot-extruded compositions gave MAE = 59.59 MPa and mean signed error = +59.59 MPa. Corresponding Ridge MAEs were 175.36 and 83.39 MPa and mean-

only reference MAEs were 140.10 and 64.95 MPa. Thus, the aggregate Cai error combines sharply different subgroup patterns, including opposite signs of prediction bias. However, chemistry and processing route also differ between these small groups, and the contrasts do not identify a causal processing effect.

For Bergmann Kremer, Extra Trees MAE was 59.40 MPa in nine sand-mold configurations and 45.02 MPa in nine metal-mold configurations; the corresponding Ridge values were 51.49 and 34.54 MPa. These condition-level summaries reflect only three nominal chemical compositions. In a separate Bergmann Kremer LOCO comparison, Extra Trees MAE was 24.13 MPa with chemistry alone and 23.65 MPa with chemistry and process features (improvement 0.48 MPa, approximately 2.0%). Across held-out AlSi9Cu3, AlSi10 and AlSi12, the respective MAEs changed from 28.59 to 28.59, 11.58 to 8.89, and 32.23 to 33.48 MPa. Thus, the small aggregate improvement was not uniform. Ridge MAE was 231.16 MPa under both representations, driven largely by instability for held-out AlSi9Cu3 (MAE 631.07 MPa). An internal training-mean comparator achieved 24.35 MPa. Because these models were trained within Kremer on only two distinct compositions in each fold, the results do not demonstrate a generalizable process-feature advantage and are not comparable to the original Zhao→Kremer errors as if they were the same experiment (Appendix Table A5).

3.2.4. Sensitivity to Zero-Encoded Unreported Elements

Fixed-model perturbation tests revealed non-negligible dependence on the assumption that source-unreported elements have zero model concentration. Across individually perturbed elements at 0.10 wt.%, the maximum absolute prediction changes for Ridge/Extra Trees were 88.21/26.35 MPa (Cai), 11.26/24.01 MPa (Kremer) and 86.89/15.45 MPa (AlSi17). Under the hypothetical joint addition of 0.10 wt.% distributed equally across the unreported elements, the Cai MAE changed from 152.37 to 160.79 MPa for Ridge and from 150.56 to 160.28 MPa for Extra Trees. The corresponding Kremer changes were 43.01→43.02 and 52.21→46.19 MPa; for AlSi17, they were 52.44→60.20 and 72.79→76.11 MPa. Some scenarios reduced error, but this does not establish that those hypothetical contents are chemically correct. In particular, the 0.10 wt.% B perturbation exceeded the maximum B concentration represented in the screened Zhao training data. These results measure model response to reconstruction choices, not the real error from unknown concentrations (Appendix Table A3).

3.3. Research Question 3: Calibration Transfer and Selective Prediction

The condition-level interval coverage and risk-filter acceptance results are summarized in Table 5.

Table 5. Median coverage of nominal 90% prediction intervals and accepted observation counts under risk thresholds set at the 80th percentile of Zhao calibration groups. The Kremer data comprise 18 processing configurations but only three nominal chemical compositions.

Exploratory Transfer Metric	Cai	Bergmann Kremer
Nominal prediction-interval coverage	90%	90%
Median empirical interval coverage	16.7%	97.2%
Median chemical-distance acceptance (80th percentile)	11/12	18/18
Median tree-spread acceptance (80th percentile)	2/12	0/18
Median joint-filter acceptance (80th percentile)	2/12	0/18
Splits with no tree-spread acceptance	3/12	9/12

To examine transferability of uncertainty estimates, twelve Zhao train–calibration splits were generated while grouping identical chemical-composition vectors. For each split, nominal 90% prediction intervals were constructed according to the group-based

procedure described in Section 2.4. Their empirical coverage was then evaluated against the separately published measurements from Cai and Bergmann Kremer. Coverage and precision of intervals. For Cai, comprising 12 different reported compositions, the median empirical coverage was 16.7%. Thus, the Zhao-calibrated prediction intervals failed to include most of the observed UTS values in this external source, revealing a substantial limitation of cross-source calibration transfer. The reported 16.7% is a median across overlapping Zhao calibration splits evaluated against the same twelve Cai observations, not a confidence-bound estimate of prediction-interval coverage for future publications. For Bergmann Kremer, which comprised 18 processing configurations represented by three nominal chemical compositions, the median empirical coverage was 97.2%. This result must be considered together with interval precision. The median full width across repeated splits was 178.18 MPa, whereas the full range of reported Kremer UTS measurements was 133.6–215.1 MPa, a span of 81.5 MPa. High observed coverage therefore did not automatically imply sufficient precision to support strength-related decisions. Because the 18 reported conditions correspond to only three nominal chemistry vectors, the 97.2% figure is a condition-weighted descriptive summary and not an independently estimated coverage probability for new alloy compositions. The original per-split results at the level of the three nominal chemical groups are presented in Table 6.

A group-level review of the original archived V90 predictions shows that all three nominal Bergmann Kremer chemistry groups were fully covered in six of the twelve Zhao splits, two groups in four splits, and only one group in two splits (Table 6). Each chemical input corresponded to six different as-cast process configurations. Thus, the median configuration-level coverage of 97.2% did not imply consistently complete coverage of chemistry groups. Because only three nominal compositions were available, and the same external observations were reused in all twelve splits, these figures are descriptive rather than a generalizable coverage guarantee.

Table 6. Original Zhao-based split-conformal coverage for each of the three Bergmann Kremer nominal alloy-composition groups. Each group contains six processing conditions; cells give covered conditions out of six. A chemistry group is fully covered only when all six conditions are within the same prediction interval. The twelve splits reuse the same external observations and are not independent replications.

Zhao Split	AlSi9Cu3	AlSi10	AlSi12	Fully Covered Groups
9001	6/6	6/6	5/6	2/3
9002	6/6	5/6	1/6	1/3
9003	6/6	6/6	3/6	2/3
9004	6/6	6/6	6/6	3/3
9005	6/6	6/6	6/6	3/3
9006	6/6	5/6	3/6	1/3
9007	6/6	6/6	3/6	2/3
9008	6/6	6/6	3/6	2/3
9009	6/6	6/6	6/6	3/3
9010	6/6	6/6	6/6	3/3
9011	6/6	6/6	6/6	3/3
9012	6/6	6/6	6/6	3/3

The findings illustrate two different limitations of chemistry-only prediction: insufficient coverage of measured responses and excessively wide intervals despite high coverage. Calibration using chemical-composition groups requires appropriate exchangeability assumptions for calibration and test units. These assumptions were not established after transfer to different publications, manufacturing processes and material states. Moreover, the same external observations were assessed across twelve overlapping Zhao splits. The

medians are descriptive summaries, not results from twelve independent external validation studies. With respect to selective acceptance of predictions, independently of prediction intervals, heuristic risk filters were assessed as a means of limiting the number of predictions retained for further evaluation. At thresholds equal to the 80th percentile of the corresponding Zhao calibration-group score, the chemical-distance filter accepted a median 11 of 12 Cai observations and all 18 Kremer configurations. The spread-of-trees filter accepted two of 12 and 0 of 18, respectively. A joint rule requiring both criteria accepted the same median numbers as the tree-spread filter. These results show that different heuristic measures can lead to radically different acceptance decisions. A very small number of accepted predictions under the tree-spread rule is not, by itself, proof of improved safety or effectiveness. When a filter accepted no observation, conditional MAE was undefined. Furthermore, the numbers of accepted Kremer configurations are not counts of chemically independent inputs.

Sensitivity to the calibration threshold. When the calibration threshold increased from the 80th to the 95th percentile, median acceptance by the tree-spread filter rose to 7/12 for Cai and 18/18 for Kremer. The joint filter accepted 6/12 and 18/18, respectively. These sensitivity results are designated for Table A2. With the tree-spread filter at the 95th percentile, median MAE among accepted predictions was 142.66 MPa for Cai and 50.46 MPa for Kremer. The respective median unfiltered MAEs in the same repeated-split experiment were 149.94 and 47.92 MPa. These are separate medians of per-split summaries and must not be combined as if they described one representative split or an independently confirmed filter improvement. Filtering did not consistently lower errors in both sources. Conditional MAE must be considered together with the number and identity of accepted configurations and the abstention rate; changes in the retained subset alone do not demonstrate lower operational risk. The analysis does not justify interpreting chemical distance or spread among tree predictions as a calibrated error probability under dataset shift. Nor has any investigated threshold been validated as an operational rule for withholding a prediction. Such application would require an explicit acceptable decision risk, acceptance rules defined in advance and assessment on previously unused casting data.

Internal Risk Ranking and Additional Robustness Checks

Risk ordering based on AURC. Across twelve Zhao splits grouped by identical chemistry vectors, the ability of local total variation (LTV) to rank distinct held-out compositions by increasing prediction error was evaluated. The discrete group-balanced AURC was calculated as defined in Section 2.4 and expressed in MPa, assigning equal weight to each distinct test composition regardless of its number of records. Median AURCs were 14.86, 14.76 and 14.58 MPa for LTV at $k = 4$, 8 and 12, respectively, compared with 18.17 MPa for the standard deviation of individual Extra Trees predictions (Figure 4). In paired comparisons within the same splits, LTV had a lower AURC in 9/12, 10/12 and 10/12 cases. Median paired AURC differences (LTV minus tree spread) were −1.08, −2.40 and −2.12 MPa, respectively. These descriptive comparisons reuse one dataset and do not represent statistically independent confirmation of the new score's superiority.

Error at a fixed acceptance level. More favorable AURC ordering did not consistently translate into lower error when 80% of distinct test compositions were retained. Median MAE_{80} values for LTV were 20.02, 19.75 and 21.41 MPa for $k = 4$, 8 and 12, respectively, compared with 19.51 MPa for across-tree spread. In paired comparisons, LTV achieved lower MAE_{80} in only six of twelve splits at each neighborhood size. MAE_{80} denotes the mean group error for the first $\lceil 0.8G \rceil$ chemically distinct test groups ranked in ascending estimated risk, where G is the number of distinct held-out groups. It is not equivalent to the 80th-percentile calibration threshold used in the cross-source selective-prediction analysis

in Section 3.3. An improved aggregated risk ranking need not imply improvement at a specific operating point.

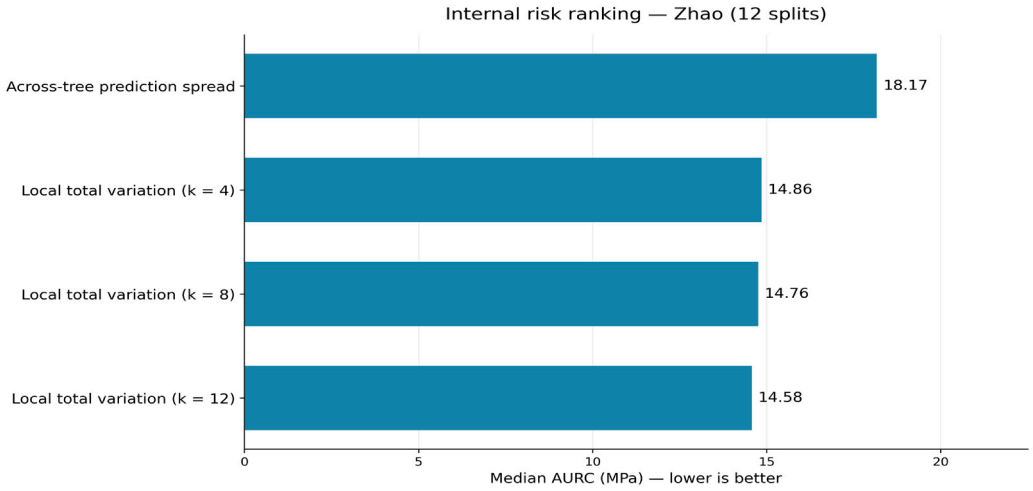

Figure 4. Zhao-only internal risk ranking across 12 composition-grouped train–test splits. AURC is the discrete, group-balanced area under the risk–coverage curve; lower values indicate more favorable ordering. LTV neighborhood sizes were $k = 4$, 8 and 12. These exploratory comparisons do not establish external validity or a lower error at fixed 80% group acceptance.

The sensitivity checks concerning Zhao data-quality screening and cross-source composition-reconstruction assumptions are summarized in Table 7.

Table 7. Sensitivity checks for training-data screening and chemical-composition reconstruction assumptions. The 211-record variant remains the primary reported analysis.

Check/Assumption	Observed Effect	Interpretation Limit
Training on 216 versus 211 Zhao records	Extra Trees MAE: Cai 150.19 vs. 150.56 MPa; Kremer 39.60 vs. 52.21 MPa	Quality screening affects Kremer; source data were not corrected
Exclusion of unusual Cai D3 input	MAE 154.80 MPa; $n = 11$	Sensitivity analysis only, not a confirmed correction
Repeated nominal Kremer chemistry	18 conditions represented by 3 vectors	No melt-specific chemistry or encoded process label
Local $k = 4$, 8 and 12	Lower AURC, unstable MAE at 80% acceptance	Exploratory parameter choice, not a validated algorithm

A supplementary check using only locally weighted within-group UTS variance, without the between-group variation of nearby means, yielded lower AURC than tree spread in 6/12, 6/12 and 7/12 splits at $k = 4$, 8 and 12, respectively. The stronger LTV ordering may therefore also be associated with variation in mean strength among neighboring compositions, rather than only the dispersion of repeated measurements at identical chemistry. The available data cannot attribute this variation to particular casting-process factors. A separate sensitivity analysis reinstated five Zhao records flagged because reported yield strength exceeded UTS. A controlled post-review extension of the screening sensitivity compared all three fixed models on both the 211-record screened and 216-record complete Zhao sets (Table 8). For Cai, the 211→216 MAEs (MPa) were 121.31→123.34 for the mean-only baseline, 152.37→155.53 for Ridge, and 150.56→150.19 for Extra Trees. For Bergmann Kremer, they were 87.34→83.29, 43.01→37.69 and 52.21→39.60 MPa, respectively. The within-source rankings were unchanged across the two training-set definitions, but the Ridge–Extra Trees advantage for Bergmann Kremer narrowed from 9.20 to 1.91 MPa.

This is a comparison of two predefined screening variants, not a demonstration of global ranking robustness. These findings do not demonstrate that the flagged UTS values were incorrect or that the 216-record variant would outperform in independent deployment. Removing the unusual Cai D3 input in an auxiliary check yielded MAE = 154.80 MPa for the remaining 11 observations. This is a sensitivity analysis of the source composition, not an independently established reason to delete D3 from the primary comparison.

Table 8. Sensitivity of the three fixed composition-only models to Zhao quality screening. MAE (MPa) for external published observations; lower values indicate lower error. For Cai, $n = 12$ reported compositions; for Bergmann Kremer, $n = 18$ processing configurations from three nominal compositions. The evaluation targets and model configurations were unchanged.

External Source	Model	MAE: Zhao $n = 211$	MAE: Zhao $n = 216$
Cai	Mean-only	121.31	123.34
Cai	Ridge	152.37	155.53
Cai	Extra Trees	150.56	150.19
Bergmann Kremer	Mean-only	87.34	83.29
Bergmann Kremer	Ridge	43.01	37.69
Bergmann Kremer	Extra Trees	52.21	39.60

All risk-ranking comparisons used twelve overlapping resamples of the same Zhao dataset. Pairing the comparisons reduces variation due to different splits but does not provide twelve independent experimental replications. The evaluation of $k = 4$, 8 and 12 was exploratory and did not independently select an optimal neighborhood size. LTV was not assessed on an independent external source as an operational abstention rule. Any apparent advantage of one neighborhood size after comparing multiple candidates on these overlapping splits requires evaluation on genuinely unused data before it can be treated as a selected operating setting. These findings suggest that local variability of training-set material responses may carry information useful for ordering predictive risk. However, they do not establish LTV as a validated, universally superior uncertainty measure or an implementation-ready decision rule for industrial applications.

The model ordering remained mean-only < Extra Trees < Ridge for Cai and Ridge < Extra Trees < mean-only for Bergmann Kremer in both screening variants. These rankings refer only to the selected datasets, and their margins changed appreciably. In particular, each Bergmann Kremer ranking is based on three nominal chemical inputs represented by six reported processing conditions each.

Sensitivity to composition-group weighting. The additional analysis compared three training strategies defined in Section 2.2 (Table 9). For Extra Trees, median row-level MAE across the twelve original Zhao grouped 80:20 splits was 26.98 MPa for strategy A, 25.75 MPa for group-mean training (B) and 25.54 MPa for inverse-group-size weighting (C). The corresponding medians were 43.71, 44.58 and 44.37 MPa for Ridge, and 66.08, 66.28 and 66.28 MPa for the mean-only predictor. These within-source comparisons are descriptive sensitivity checks based on repeated use of the same Zhao data, not independent external validation.

The weighting strategy altered prediction errors in both external sources. For Cai, Extra Trees MAE increased from 150.56 MPa (A) to 153.79 MPa (B) and 157.61 MPa (C), whereas for Bergmann Kremer, it changed from 52.21 MPa (A) to 47.02 MPa (B) and 50.15 MPa (C). The relative ranking of Ridge and Extra Trees for Cai changed under B and C, although the mean-only predictor remained the lowest-MAE model for Cai and Ridge remained the lowest-MAE model for Bergmann Kremer under all three strategies. Thus, the detailed model ordering was not invariant to training-group weighting; the broader observation that predictive performance depends strongly on the external source remained. Each of the

three Bergmann Kremer nominal compositions had exactly six processing configurations. Consequently, equal weighting of the three compositions produces the same overall MAE as equal weighting of the eighteen conditions; this does not make the processing configurations chemically independent, and it differs from computing an error against a group-mean UTS.

Table 9. Sensitivity of MAE (MPa) to the representation and weighting of repeated Zhao chemical-composition groups. A: 211 equally weighted records; B: 151 group-mean training targets; C: 211 records with inverse composition-group-size weights. Models retained the original hyperparameters. Lower MAE indicates lower error.

External Source	Model	A: Equal Rows	B: Group Mean	C: Inverse Size
Cai	Mean-only	121.31	119.18	119.18
Cai	Ridge	152.37	153.52	154.39
Cai	Extra Trees	150.56	153.79	157.61
Bergmann Kremer	Mean-only	87.34	91.61	91.61
Bergmann Kremer	Ridge	43.01	45.60	44.47
Bergmann Kremer	Extra Trees	52.21	47.02	50.15

Research-question synthesis and robustness. For RQ1, the repeated-input ambiguity remains a directly calculable descriptive property of the Zhao table (minimum fitted RMSE 23.07 MPa for 216 rows and 23.34 MPa after screening to 211), rather than a prospective performance guarantee. For RQ2, the importance of source-dependent errors and the identity of the best baseline for Cai versus Ridge for Kremer survived the investigated row-screening and chemistry-group-weighting variants, while absolute errors and some model-to-model rankings depended on the weighting strategy; reconstructed inputs additionally showed potentially substantial what-if sensitivity. The process-stratified and three-composition LOCO exercises added descriptive context but did not isolate processing effects or establish improved industrial transfer. For RQ3, poor interval transfer to Cai, very wide intervals despite high coverage for Kremer, and the disconnect between AURC ordering and fixed-acceptance MAE remained the principal findings. These are source-specific conclusions under the analyses undertaken, not universal statements concerning aluminum foundries.

4. Discussion

4.1. Information Sufficiency Versus Algorithmic Complexity

The results show that limitations in the reliability of UTS prediction for Al–Si alloys arise not only from the choice of regression algorithm but also from the information contained in the input representation. Any deterministic model receiving the same 17-element chemical-composition vector must assign the same prediction to records with an identical recorded input. If those observations have different UTS values, increasing model complexity alone cannot reproduce these differences. In the Zhao table analyzed here, 216 rows represented 156 unique chemical-composition vectors, and 92 observations belonged to repeated-input groups. The empirical lower RMSE bound of 23.07 MPa for the full dataset represents the smallest fitted error achievable when all observations with identical inputs receive one constant prediction. It is not a universal accuracy limit for machine learning: it applies to this particular table, the selected chemical representation and deterministic prediction, rather than to models supplied with additional information on processing and material state. Different outcomes for identically recorded compositions may reflect variation in manufacturing conditions, thermal history, microstructure and defects, as well as measurement variation or limited precision of composition reporting. The Zhao dataset does not provide sufficient metadata to attribute these causes to individual observations. Therefore, the contributions of omitted process variables, measurement uncertainty and other sources of variation cannot be quantified separately. Furthermore, the present chemistry-only analysis

is not a direct assessment of the original Zhao approach, which also incorporated a broader range of features [10].

The importance of missing variables is illustrated by Bergmann Kremer [23]. In nine paired comparisons, metal-mold configurations exhibited UTS values on average 16.94 MPa higher than their sand-mold counterparts. The reconstructed chemical inputs were identical for the two processing variants, so the model could not directly reproduce their technological contrasts. Published AlSi7Mg studies also document associations among cooling conditions, microstructure and mechanical properties [3,4]. These comparisons do not, however, isolate the causal contribution of mold type. An analogous illustration is the additional AlSi17 series [24]. Twelve mean UTS values for different overheating temperatures and two mold types were assigned one reconstructed chemical-composition vector. Extra Trees therefore produced the same prediction of 177.13 MPa for all conditions, whereas published mean strengths ranged from 198 to 294 MPa. This shows the inability to distinguish processing conditions under the available input representation but does not allow the full predictive error to be attributed solely to overheating temperature, mold type or any one factor. The findings do not demonstrate that chemical composition is an uninformative predictor of strength. Rather, its sufficiency depends on the task, the variability of manufacturing processes and the intended application. For technology-oriented decision support, additional process and material-state descriptors should be considered, but those features must genuinely be available at the time of the relevant decision. Independent, prospective validation against pre-established evaluation criteria would also be needed. Increasing algorithmic sophistication without improving the representation of relevant variables is therefore insufficient as a strategy for improving cross-source predictive reliability. The central requirement is alignment between the input information, processes represented in training and conditions under which predictions are used.

4.2. What Differences Between External Sources Do—And Do Not—Demonstrate

The comparisons demonstrated a marked dependence of chemistry-only model performance on the source to which predictions were transferred. Extra Trees trained on 211 Zhao records gave MAE = 150.56 MPa for Cai and 52.21 MPa for Bergmann Kremer. The mean signed error also changed direction: the Cai measurements were predominantly underpredicted, whereas the Kremer measurements were overpredicted. Thus, neither a common error magnitude nor a consistent direction of bias can automatically be assumed for new materials datasets. The relative performance of the evaluated methods was source-dependent as well. On Cai, a Zhao-trained mean predictor achieved a lower MAE than Extra Trees or Ridge. On Kremer, Ridge achieved the lowest MAE. This observation does not confirm universal superiority of any particular algorithm; rather, predictive complexity and success within one data region do not suffice to predict performance in a different materials-processing domain. The evaluated sources also had different observation structures. Cai comprised twelve distinct reported chemical compositions and a range of processing states. By contrast, eighteen Kremer processing configurations represented only three nominal compositions. Observations produced under different technological conditions are not automatically independent with respect to the chemical information supplied to the model. Error values computed for these sources should therefore not be treated as directly comparable estimates obtained under identical experimental conditions. Different composition ranges, manufacturing and heat-treatment conditions, microstructural states, UTS distributions and assumptions used to reconstruct the chemical inputs may all have affected cross-source performance. In particular, assigning zero to unreported elements is a computational convention, not experimental evidence that those elements are absent. The available data cannot separately quantify the contribution of each factor to prediction

error. Hence, observed differences are not quantitative estimates of casting-process effects, chemical covariate shift alone or one specific failure mechanism of generalization.

It is also important to distinguish independence of the data used for model fitting from independence of the overall evaluation procedure. UTS measurements from Cai and Bergmann Kremer were not used to train the principal Zhao model. Nevertheless, these publications were known and examined during project development. The comparisons are therefore exploratory cross-source assessments, not fully independent, prospectively blinded external validation under a protocol fixed before the published outcomes were inspected. The additional AlSi17 series [24] illustrates limits of composition-only representation in another metallurgical experiment. Because it comes from a publication by a coauthor and its outcomes were available during preparation, it must not be presented as a fully independent, blinded test. Furthermore, its twelve mean UTS values corresponded to one reconstructed chemical input and differing processing conditions rather than twelve independent compositions. The results neither undermine the general usefulness of machine learning for alloy-property prediction nor refute models that incorporate information beyond chemical composition. They are also not a direct evaluation of the original Zhao method [10]. They are consistent with broader problems of realistic generalization and domain adaptation discussed in the materials informatics literature [12–14]. The contribution here is a quantitative demonstration for the specific Al–Si sources examined, rather than a new theory of model generalization. Convincing evidence of transferability would require a prespecified assessment protocol, clearly identified independent experimental units and measurements from new, previously unused materials series. Only such validation could determine whether the patterns reported here persist beyond the published datasets analyzed in this study.

4.3. Reliability Screening Cannot Replace a Domain-Appropriate Model

The analyses show that calibrating uncertainty on a source dataset does not, by itself, ensure reliable predictions after transfer to different materials and processing conditions. Nominal 90% intervals calibrated on Zhao achieved median coverages of 16.7% for Cai and 97.2% for Bergmann Kremer. At the same time, their median full width was 178.18 MPa, exceeding the 81.5 MPa span of the reported Kremer UTS measurements. These results illustrate two different limitations: inadequate coverage of observed strengths and inadequate precision despite high coverage. Either can reduce the value of a model for technological decisions. An interval that misses the measured UTS may lead to incorrect assessment; a very wide interval may not allow a definite conclusion about whether an alloy meets mechanical requirements. Coverage alone is therefore insufficient for assessing industrial usefulness. Heuristic risk filters did not consistently remove these limitations. Rules based on chemical distance and prediction spread among individual trees led to different external acceptance fractions, with marked dependence on whether the 80th or 95th calibration percentile was selected. Raising the tree-spread threshold increased acceptance but did not improve prediction errors consistently in both sources. None of these heuristics should be interpreted as a calibrated probability of error or a validated rule for approving technological predictions.

The internal risk-ranking study showed that LTV may contain information useful for ordering chemical compositions by their prediction error. For $k = 12$, median AURC was 14.58 MPa compared with 18.17 MPa for tree-prediction spread. However, this advantage was not consistently retained at a fixed acceptance of 80% of different composition groups: median MAE_{80} values were 21.41 and 19.51 MPa, respectively. The analysis used repeated splits of Zhao alone and did not include independent external validation of LTV. Future industrial applications should distinguish three levels of assessment: regression-model accuracy, correctness of uncertainty calibration and effectiveness of the decision rule. Each

must be examined for a defined casting technology, chemical range, material condition and strength requirements. Recalibrating intervals on representative target-domain observations is a possible research direction but does not remove missing technological features or automatically improve point predictions of UTS. Before a model can be regarded as a decision-support tool, acceptable error types and consequences, abstention rules and material acceptance criteria must be defined in advance. Effectiveness should then be assessed using genuinely new, independent experimental series. The present findings motivate this requirement, but do not demonstrate a validated operational decision-support system.

4.4. Scientific Contribution in Relation to Previous Research

The original contribution goes beyond the unsurprising fact that identical chemistry inputs cannot distinguish different processing states. It measures three separable consequences: an empirical input–UTS ambiguity (minimum chemistry-only fitting RMSE of 23.07 MPa in the complete Zhao table); source-specific error magnitude, bias and sensitivity to group weighting, data screening and composition reconstruction; and imperfect transfer of uncertainty calibration and selective prediction. In addition, process-stratified errors and the composition-held-out Kremer exercise quantify both the potential importance of missing state descriptors and the present lack of evidence that adding those descriptors alone improves prediction on unseen alloys. The findings remained qualitatively robust in the restricted senses specified above; neither numerical invariance under all assumptions nor broad industrial generalization is claimed.

Previous studies have demonstrated the potential of machine-learning methods for predicting the mechanical properties of aluminum alloys, including Al–Si compositions [7–10]. Other research has addressed data leakage, out-of-distribution generalization and realistic model evaluation in materials informatics [11–14]. Prediction-interval calibration and selective prediction have also been extensively investigated in the broader machine-learning literature [15–18]. The present study does not introduce a new regression algorithm, conformal prediction method or general theory of distribution shift. Its contribution is the quantitative examination of how these established methodological concerns manifest in chemical-composition-based UTS prediction for specific published Al–Si alloy datasets.

First, the analysis identifies and quantifies ambiguity in the recorded composition–strength relationship. The Zhao dataset contained 216 observations corresponding to 156 unique chemical-composition vectors, with 92 observations distributed across 32 repeated-input groups. The minimum in-sample RMSE attainable by assigning one deterministic fitted value to each distinct recorded composition was 23.07 MPa. This result quantifies the information limitation of the particular composition-only representation, without implying a universal error bound for models incorporating processing or microstructural information.

Second, the cross-source comparisons reveal substantial differences in predictive performance. The Extra Trees model trained on 211 screened Zhao observations yielded MAEs of 150.56 MPa for Cai and 52.21 MPa for Bergmann Kremer, with opposite directions of prediction bias. The mean-only baseline outperformed Extra Trees for Cai, whereas Ridge achieved the lowest MAE for Bergmann Kremer. These findings provide a quantitative, alloy-specific illustration of source-dependent model reliability. They do not isolate the individual effects of composition, processing conditions or measurement uncertainty, and they should not be interpreted as a direct assessment of the broader Zhao modeling approach.

Third, the study demonstrates that uncertainty calibration and practical prediction precision must be evaluated separately. Nominal 90% prediction intervals calibrated using the Zhao dataset achieved median condition-level coverages of 16.7% for Cai and 97.2% for Bergmann Kremer, while the median full interval width reached 178.18 MPa. The

additional group-level assessment of Bergmann Kremer further clarifies that high coverage across processing configurations does not guarantee complete coverage of chemically distinct groups.

Fourth, the internal risk-ranking analysis shows that a favorable aggregate ranking metric does not necessarily translate into improved performance at a fixed acceptance level. For example, LTV at $k = 12$ yielded a median AURC of 14.58 MPa compared with 18.17 MPa for tree-prediction spread but did not outperform that reference at 80% group acceptance. This distinction limits claims concerning the operational effectiveness of heuristic rejection rules.

Taken together, these results establish a quantitative framework for distinguishing input-information limitations, cross-source predictive errors, uncertainty-calibration performance and decision-level reliability in Al–Si alloy modeling. The Mean-only, Ridge and Extra Trees comparison illustrates source- and representation-dependent reliability, not universal algorithmic superiority. Its scientific originality lies in the integrated assessment of these limitations and their metallurgical implications.

The relevance to sustainable manufacturing is conditional: reliable predictions and appropriately validated decision rules are prerequisites for assessing whether AI-supported materials decisions can reduce unnecessary testing, rejected castings, remelting or resource consumption. None of these environmental benefits was directly measured in the present study, and their demonstration requires prospective industrial evaluation.

4.5. Implications for Sustainable Manufacturing and AI-Assisted Decision-Making

The principal relevance of these results to sustainable manufacturing is the identification of prerequisites that should be fulfilled before AI predictions are applied to materials and processing decisions. The analysis identifies predictive-reliability requirements for future AI-assisted manufacturing decisions, but does not demonstrate reductions in energy use, material losses, rejected castings or greenhouse-gas emissions. Rather, they show that insufficient predictive reliability may limit the value of such models for manufacturing decisions. The different Extra Trees errors in the exploratory cross-source assessments (MAE = 150.56 MPa for Cai and 52.21 MPa for Bergmann Kremer) illustrate why performance reported for one materials-data source cannot be assumed for another; these errors do not quantify incorrect production decisions. A possible use of a chemistry-only model is preliminary comparison of Al–Si compositions under previously specified manufacturing conditions and material states. Such a model cannot independently select mold type, overheating temperature or heat-treatment conditions: configurations differing only in these parameters may receive identical predictions because their encoded chemical inputs are the same. The separate coauthor-published AlSi17 processing series further illustrates this limit: one reconstructed chemistry input led to a constant prediction of 177.13 MPa for all 12 processing conditions, whereas the published condition means ranged from 198 to 294 MPa. These previously available observations do not constitute independent prospective validation. In the case of secondary-aluminum feedstocks, even preliminary composition screening would require verification for the actual recycled-alloy chemistry range and casting route; the present study did not test recycled-alloy qualification as an industrial decision task. In foundry practice, two kinds of decision error are particularly relevant. UTS underprediction could lead to rejection of a variant that meets mechanical requirements, or to additional qualification testing. Overprediction could encourage premature acceptance of a material that fails those requirements. Actual consequences depend on threshold properties, acceptance procedures and subsequent quality control. MAE or RMSE alone cannot establish the number of erroneous decisions or their environmental costs. Lower average prediction error, by itself, would not establish fewer incorrect acceptances or rejections at a specified UTS threshold, still less a reduction in resource use.

A prospective decision pathway could use a composition-only prediction to prioritize Al–Si candidates for confirmatory mechanical testing, provided the model had first been independently validated for the relevant chemistry range, casting technology and material state. A predicted UTS, even when accompanied by a prediction interval, would not itself demonstrate compliance with a specified mechanical-strength requirement. If the input lies outside the validated domain, the uncertainty is excessive, or the prediction interval does not permit a sufficiently reliable assessment relative to the UTS requirement, the AI prediction should be withheld and the established qualification procedure followed. These are proposed decision principles rather than validated industrial thresholds: the present study did not establish a reliable operational abstention rule.

Future industrial assessment should therefore begin by defining the particular decision task, including required UTS, acceptable risks of incorrect approval and rejection, conditions for withholding a prediction and the required extent of confirmatory testing. Performance of an AI-supported procedure should be compared with the actual reference workflow without the model. Validation should use new casting batches that were not used previously to train models, choose parameters or set decision thresholds. Only such a prospective assessment could link predictive quality to sustainability indicators. Suitable measures may include the mass and fraction of rejected castings, metal requiring remelting, charge-material consumption, the number of additional qualification tests, energy per unit of accepted output and associated greenhouse-gas emissions. Comparisons require common system boundaries and a defined functional unit, such as 1 kg of accepted castings meeting specified quality requirements. A broader environmental assessment could subsequently include an appropriately scoped life-cycle analysis. Studies on circular-economy alloy selection, melt preparation for recycled aluminum and environmental impacts of heat-treatment processes demonstrate the relevance of metallurgical decisions to resource utilization [19–21]. However, their findings do not establish that the models examined here would achieve comparable environmental effects. Their role here is to motivate which production outcomes would need to be monitored in a future implementation, not to provide evidence that composition-only AI is already beneficial for recycled-metal processing.

The proposed framework for assessing prospective AI-assisted foundry decisions is summarized in Table 10.

Table 10 presents a proposed framework for evaluating prospective industrial use of AI. It is a conceptual protocol for future validation, not an inventory of benefits achieved in the present study. In particular, recommending different casting or heat-treatment conditions would require an expanded model with appropriate process descriptors and independent validation. A prospective assessment should distinguish three successive levels of evidence: (i) independent validation of tensile-strength predictions and uncertainty estimates in the intended casting domain; (ii) evaluation of predefined accept, reject and abstain decisions against the existing quality-control workflow; and (iii) measurement of resulting production and environmental indicators under consistent system boundaries. The present exploratory comparisons motivate these requirements but do not provide independent prospective target-domain validation, an evaluation of operational accept/reject decisions, or measurements of manufacturing and environmental outcomes. Ultimately, the results support demonstrating predictive reliability for the predefined application domain as a prerequisite for responsible AI use in foundries. They do not demonstrate that the evaluated models reduce resource consumption or manufacturing environmental impact. Establishing such benefits requires separate prospective investigations of decision quality and measurable production and environmental outcomes.

Table 10. Proposed evidence framework for assessing AI-assisted foundry decisions. This is a conceptual evaluation protocol, not a measured environmental benefit.

Potential Application of AI Output	Demonstrated or Plausible Error Pathway	Evidence Required Before Claiming Environmental Benefits
Preliminary alloy candidate screening before qualification	Source-specific underprediction may reject an appropriate alloy; overprediction may favor an unsuitable material	Record false-acceptance/rejection rates on previously unused castings and the number of confirmatory tests
Comparing casting methods or heat-treatment states in a future process-aware model	A chemistry-only model assigns identical predictions to conditions sharing one recorded chemical input	Validate a model incorporating processing descriptors on new casting series under prespecified criteria
Claiming reductions in rejects, energy use or remelting	Prediction errors may affect foundry decisions, but the direction and magnitude of resource impacts are unknown	Measure rejected casting mass, remelted metal, electricity/fuel consumption and emissions inventories within defined system boundaries
Withholding a prediction when predefined reliability criteria are not met	A model applied outside a verified chemistry/process domain, or with excessive uncertainty, can support an incorrect material-qualification decision	Prospectively assess abstention criteria, confirmation-testing needs and decision errors versus conventional qualification before evaluating any environmental effect

5. Limitations

This study has limitations related to the available data, the chosen chemistry representation, model validation and the interpretation of its findings.

First, the publicly released Zhao dataset lacks identifiers linking individual records to source experiments, complete processing histories and information supporting independent grouping by materials batch or publication. Grouping identical recorded chemical compositions reduces direct duplicate-input leakage between training and test sets but does not establish experimental independence. Overlap between unidentified original Zhao sources and the publications used for external comparison also cannot be excluded.

Second, the cross-source assessment was limited to Cai, with 12 distinct reported compositions, and Bergmann Kremer, whose 18 processing configurations represented only three nominal chemistry inputs. The separate AlSi17 illustration from a coauthor publication [24] comprised twelve condition-level mean UTS values assigned to one reconstructed chemical input. External UTS values were not used to train the principal model, but these publications were known during study development. The analyses are exploratory cross-source comparisons rather than prospective, blinded validation. The reported external errors are descriptive and do not support population-level confidence intervals for future foundries.

Third, input preparation and data-quality screening introduced assumptions. Conversion from reported mass to atomic fractions required a convention for elements not quantified in the source publications. Assigning zero does not prove their absence. The AlSi17 model input also omitted the reported phosphorus addition because P was not a feature. Five Zhao rows with yield strength exceeding UTS were excluded as quality flags, not as proof that their UTS was wrong. The 211-versus-216 screening comparison preserved the model order in the two examined external sources but changed performance margins markedly, particularly for Bergmann Kremer (Ridge versus Extra Trees: 9.20 versus 1.91 MPa). Accordingly, neither the magnitude of model differences nor their wider generalizability is established as stable across dataset definitions. For Cai, tabulated alloy-design

compositions should not be conflated with independent complete chemical analyses of every tested specimen.

Unequal representation of identical Zhao chemical inputs is another modeling choice: the original models treated each retained record as a training observation. The post-review comparison of group-mean targets and inverse-group-size weights showed changes in cross-source errors, including a reversal of the Ridge–Extra Trees ordering for Cai under the alternative strategies. These analyses do not establish an optimal weighting rule for other published datasets or independent foundry deployment.

Fourth, uncertainty calibration and selective prediction were evaluated using repeated splits of one Zhao dataset, not independent experimental replications. The assumptions supporting nominal prediction-interval coverage were not established under cross-source shifts. Moreover, lower AURC values for LTV did not consistently yield lower error at fixed 80% acceptance of distinct compositions. No investigated heuristic was validated externally as an operational decision rule.

Fifth, computational reproducibility and data availability impose constraints. Re-execution of fixed scripts reproduced archived results for specified inputs and implementations, but independent replication of the complete analysis requires source data, code and unambiguous transcription and chemistry-reconstruction instructions. The importance of these requirements is emphasized in [25]. The computational package, including the additional post-review weighting sensitivity, is supplied for peer review, and the author-developed materials can be made available by the corresponding author upon reasonable request. The original third-party observation-level records are not redistributed in this package. Reuse permissions for source-derived data must be assessed separately; reference [25] does not determine those legal conditions.

Finally, these analyses cannot quantitatively separate the causal effects of composition, processing parameters, microstructure and measurement variability on prediction errors. No real production decisions were evaluated, and no energy consumption, material losses, casting rejects or greenhouse-gas emissions were measured. The study therefore identifies limitations and preconditions for future AI use in foundries without demonstrating actual environmental benefits.

The Zhao source also includes six compositions with Si recorded as zero, and the published table lacks row-linked processing metadata; the interpretation is therefore limited to the heterogeneous aluminum-rich training observations actually used. The zero-assignment sensitivity tests vary hypothetical rather than measured concentrations. The process-aware LOCO analysis involves only three nominal compositions and is not a prospective validation of either a process-aware Zhao model or industrial predictive reliability.

6. Conclusions

The three research questions lead to the following five conclusions. These findings concern the source datasets and evaluation procedures examined; environmental performance was not measured.

1. RQ1—Identical recorded chemical inputs were associated with different measured UTS values. The Zhao source comprised 216 rows but 156 unique chemistry vectors, including 92 records in 32 repeated groups. The descriptive minimum chemistry-only in-sample RMSE was 23.07 MPa (23.34 MPa for 211 screened rows). These are in-sample fitting bounds for the Zhao table and its recorded chemical-input representation, not universal irreducible errors or limits on prospective predictive accuracy.
2. RQ2—Cross-source prediction performance and signed bias depended strongly on the selected publication and material conditions. For Zhao-trained Extra Trees, MAE was 150.56 MPa for Cai and 52.21 MPa for Bergmann Kremer; the strongest corresponding

references were the training mean (121.31 MPa) and Ridge (43.01 MPa), respectively. These leading source-specific models remained so under the tested screening and group-weighting strategies, while error magnitudes and some nonleading model rankings changed. Hypothetical imputation perturbations also changed predictions substantially; their effects are model sensitivities, not measured chemical-analysis errors or evidence of universal algorithmic superiority.

3. RQ3—Prediction-uncertainty calibration and selective prediction did not provide uniformly reliable transfer. Nominal 90% Zhao-calibrated prediction intervals produced median coverage of 16.7% for Cai and 97.2% for Kremer with a median full width of 178.18 MPa. Although the local total-variation score achieved median AURC of 14.58 MPa versus 18.17 MPa for tree spread, it did not consistently improve error at fixed 80% composition-group acceptance. Operational rejection rules require separate validation.
4. Additional processing analyses—Cai T6 versus extruded subsets and Bergmann Kremer mold categories had markedly different descriptive errors, but chemistry and processing differences could not be separated causally. In an exploratory three-composition Kremer LOCO test, adding process descriptors reduced Extra Trees MAE by only 0.48 MPa (24.13 to 23.65 MPa) and yielded no aggregate Ridge MAE improvement. This very small and inconsistent result does not show that processing variables are unimportant, nor does it establish generalization to new industrial alloy populations.
5. Manufacturing and sustainability—The evidence supports explicit input-provenance reporting, process- and material-state descriptors, independent validation and predefined criteria for withholding unreliable AI predictions before any foundry decision support. Prospective comparison against conventional qualification is necessary before claiming changes in rejected castings, remelting, energy consumption or greenhouse-gas emissions. No environmental outcomes were quantified; the sustainability implications remain prospective rather than demonstrated benefits.

Author Contributions: Methodology, J.P. and T.M.; validation, J.P., B.O. and T.M.; formal analysis, S.B. and B.O.; investigation, T.M. and J.P.; resources, B.O. and S.B.; data curation, J.P., S.B. and T.M.; writing—original draft preparation, T.M.; writing—review and editing, J.P., B.O. and S.B.; visualization, B.O. and S.B.; supervision, T.M. and B.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Silesian University of Technology, grant number 11/020/BK_26/0152.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The computational scripts, input-data reconstruction procedures, software requirements, and derived numerical results supporting this study are available from the corresponding author upon reasonable request. The original datasets were obtained from previously published studies and their supplementary materials. Restrictions may apply to the redistribution of third-party data; therefore, the relevant source references and reconstruction procedures are provided to facilitate independent reproduction of the analyses. A computational reproducibility package has been provided with the revised submission for peer review.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Additional Results from the Original Analyses

Table A1. Archived within-source Zhao benchmark: twelve grouped held-out splits (test_size = 0.25, seeds 9001–9012) for the fixed Extra Trees model. MAE and RMSE are row-weighted; the final column gives composition-group-balanced MAE. All errors are expressed in MPa. "Median" is the median across overlapping splits, not an independent replication estimate.

Seed	Rows	Groups	MAE	RMSE	R^2	Group MAE
9001	50	38	22.45	34.27	0.800	19.00
9002	62	38	46.17	73.67	−0.098	19.72
9003	59	38	48.32	74.21	0.067	24.70
9004	67	38	52.37	77.48	−0.116	27.97
9005	44	38	23.59	34.52	0.851	23.40
9006	47	38	21.88	33.15	0.845	20.87
9007	64	38	47.40	73.11	0.083	24.67
9008	55	38	27.69	36.90	0.718	25.87
9009	50	38	29.61	42.54	0.669	28.97
9010	45	38	31.15	42.16	0.764	30.14
9011	50	38	23.09	32.99	0.769	21.80
9012	56	38	29.63	43.64	0.667	27.76
MEDIAN	—	—	29.62	42.35	0.694	24.69

Table A2 reports the archived 95th-percentile calibration-threshold sensitivity analysis from the original twelve Zhao grouped splits. "Always" is the unfiltered comparator; "distance" uses chemical distance, "spread" the spread of individual tree predictions, and "both" their conjunction. All acceptance and MAE figures are medians across the twelve overlapping splits. The nominal interval coverage and width columns are associated with the 90% prediction intervals, not with the filter-acceptance threshold.

Table A2. Archived threshold-95% sensitivity results for acceptance filters on Cai and Bergmann Kremer; medians across twelve Zhao grouped splits. "Accepted C/G" denotes median numbers of accepted reported configurations/distinct chemistry groups; "Zero" is the number of splits with no accepted predictions. "PI coverage/width" gives median coverage and median full width (MPa) of the separate nominal 90% intervals.

Source	Rule	Accepted C/G	Accepted %	MAE All	MAE Kept	Zero	PI cov./Width
Cai	always	12/12	100.0%	149.94	149.94	0	16.7%/178.18
Cai	distance	11/11	91.7%	149.94	154.69	0	16.7%/178.18
Cai	spread	7/7	58.3%	149.94	142.66	0	16.7%/178.18
Cai	both	6/6	50.0%	149.94	149.05	0	16.7%/178.18
B. Kremer	always	18/3	100.0%	47.92	47.92	0	97.2%/178.18
B. Kremer	distance	18/3	100.0%	47.92	47.92	0	97.2%/178.18
B. Kremer	spread	18/3	100.0%	47.92	50.46	0	97.2%/178.18
B. Kremer	both	18/3	100.0%	47.92	50.46	0	97.2%/178.18

For Table A2, the calibration acceptance-score threshold was the 95th percentile for all rules; each row summarizes twelve repeated splits. Conditional MAE is calculated for retained observations where at least one prediction is accepted. Filtering and interval coverage address different questions, and the separate reported medians do not describe a single representative split.

The new appendices below report hypothetical input-reconstruction sensitivity, reconstructed external compositions and the exploratory within-Kremer LOCO comparison. No original third-party observation-level UTS tables are reproduced.

Table A3. Sensitivity of fixed Zhao-trained model predictions to hypothetical nonzero compositions of elements unreported by external sources. All errors and prediction differences are MPa. Joint scenario: total 0.10 wt.% distributed equally across unreported features; single scenario: one unreported feature changed to 0.10 wt.%. Values are not actual chemical imputation errors.

Source	Model	Original MAE	Joint 0.10 wt.% MAE	Largest |ΔUTS|, Single 0.10 wt.%
Cai	Ridge	152.37	160.79	88.21
Cai	ExtraTrees	150.56	160.28	26.35
Bergmann Kremer	Ridge	43.01	43.02	11.26
Bergmann Kremer	ExtraTrees	52.21	46.19	24.01
AlSi17	Ridge	52.44	60.20	86.89
AlSi17	ExtraTrees	72.79	76.11	15.45

The following panels list the 16 unique reconstructed external chemistry vectors (12 Cai design compositions, three nominal Bergmann Kremer compositions and one AlSi17 composition). All values are mass percentages before conversion to model atomic fractions. An asterisk (*) marks a computational zero assigned because the source did not quantify that element; unmarked 0 values were reported as zero. Superscript-style "b" after aluminum indicates balance by difference, not an independent assay. AlSi17 also had a reported 0.05 wt.% P addition, excluded from the 17-feature space.

Table A4 continues with the remaining eight model features in the same 16-source-vector order.

Table A4. (**a**). Reconstructed external chemical inputs (wt.%): Al, B, Cr, Cu, Er, Fe, Mg, Mn and Ni. Asterisk = computational zero for an unreported feature; "b" = Al balance, see accompanying text. (**b**). Reconstructed external chemical inputs (wt.%): Sc, Si, Sn, Sr, Ti, V, Zn and Zr. Data-source and imputation conventions are identical to (**a**).

(**a**)

Composition	Al	B	Cr	Cu	Er	Fe	Mg	Mn	Ni
Cai A1	92.2	0 *	0	0.2	0 *	0.1	0.3	0	0
Cai A2	91.4	0 *	0	0.2	0 *	0.1	0.3	0	0
Cai A3	90.6	0 *	0	0.2	0 *	0.1	0.3	0	0
Cai B1	91.6	0 *	0	1.7	0 *	0.1	0.2	0	0
Cai B2	93.9	0 *	0	1.5	0 *	0.1	0.3	0	0
Cai B3	92.9	0 *	0	1.6	0 *	0.1	0.3	0	0
Cai C1	90.3	0 *	0	2.9	0 *	0.2	0.3	0.1	0
Cai C2	89.3	0 *	0	3.2	0 *	0.2	0.4	0.1	0.1
Cai C3	88.5	0 *	0	3.4	0 *	0.2	0.4	0.1	0.1
Cai D1	87.1	0 *	0	0	0 *	0	0.6	0	0
Cai D2	87.7	0 *	0	0.1	0 *	0	0.6	0	0
Cai D3	77.6	0 *	0	0.1	0 *	0	11.9	0	0
Kremer AlSi9	87.2989 b	0.0041	0 *	2.14	0 *	0.692	0.315	0.402	0 *
Kremer AlSi10	89.539 b	0.003	0 *	0.01	0 *	0.156	0.308	0.206	0 *
Kremer AlSi12	87.8228 b	0.0032	0 *	0.002	0 *	0.188	0.005	0.003	0 *
AlSi17	82.87 b	0 *	0 *	0.07	0 *	0.34	0.04	0.03	0.02

Table A4. *Cont.*

(**b**)								
Composition	**Sc**	**Si**	**Sn**	**Sr**	**Ti**	**V**	**Zn**	**Zr**
Cai A1	0 *	7.2	0 *	0 *	0 *	0 *	0	0 *
Cai A2	0 *	8	0 *	0 *	0 *	0 *	0	0 *
Cai A3	0 *	8.8	0 *	0 *	0 *	0 *	0	0 *
Cai B1	0 *	6.2	0 *	0 *	0 *	0 *	0	0 *
Cai B2	0 *	4	0 *	0 *	0 *	0 *	0	0 *
Cai B3	0 *	5	0 *	0 *	0 *	0 *	0	0 *
Cai C1	0 *	6.2	0 *	0 *	0 *	0 *	0	0 *
Cai C2	0 *	6.7	0 *	0 *	0 *	0 *	0	0 *
Cai C3	0 *	7.4	0 *	0 *	0 *	0 *	0	0 *
Cai D1	0 *	12.3	0 *	0 *	0 *	0 *	0	0 *
Cai D2	0 *	11.6	0 *	0 *	0 *	0 *	0	0 *
Cai D3	0 *	10.4	0 *	0 *	0 *	0 *	0	0 *
Kremer AlSi9	0 *	8.52	0 *	0.018	0.053	0 *	0.557	0 *
Kremer AlSi10	0 *	9.62	0 *	0.036	0.111	0 *	0.011	0 *
Kremer AlSi12	0 *	11.84	0 *	0.02	0.11	0 *	0.006	0 *
AlSi17	0 *	16.63	0 *	0 *	0 *	0 *	0 *	0 *

The LOCO diagnostic used identical held-out composition groups for both representations. Numbers below are MAE (MPa), with six test conditions per alloy. The training mean reference for this particular within-Kremer exercise was 24.35 MPa.

Table A5. Exploratory Leave-One-Composition-Out errors for chemistry-only versus chemistry-plus-process models fitted within the Bergmann Kremer dataset; three nominal compositions, 18 process configurations. This is not the Zhao-to-Kremer cross-source test.

Model	**Features**	**AlSi9Cu3**	**AlSi10**	**AlSi12**	**All (18)**
ExtraTrees	17 elements	28.59	11.58	32.23	24.13
ExtraTrees	17 + process	28.59	8.89	33.48	23.65
Ridge	17 elements	631.07	46.11	16.30	231.16
Ridge	17 + process	631.07	46.11	16.30	231.16

References

1. Callegari, B.; Lima, T.N.; Coelho, R.S. The Influence of Alloying Elements on the Microstructure and Properties of Al-Si-Based Casting Alloys: A Review. *Metals* **2023**, *13*, 1174. [CrossRef]
2. Wu, D.; Wang, Y.; Li, F.; Xue, Y.; Li, J.; Su, R. Controlling eutectic Si in Al–Si casting alloys: From fundamental mechanisms to performance. *J. Mater. Res. Technol.* **2026**, *41*, 7741–7778. [CrossRef]
3. Li, L.; Li, D.; Feng, J.; Zhang, Y.; Kang, Y. Effect of Cooling Rates on the Microstructure and Mechanical Property of La Modified Al7SiMg Alloys Processed by Gravity Die Casting and Semi-Solid Die Casting. *Metals* **2020**, *10*, 549. [CrossRef]
4. Kamińska, J.; Angrecki, M.; Puzio, S. Effect of the Type of Inorganic Binder on the Microstructure and Properties of AlSi7Mg Alloy Castings Made by Ablation Casting Technology. *Materials* **2022**, *15*, 4912. [CrossRef] [PubMed]
5. Manani, S.; Pradhan, A.K. Effects of melt thermal treatment on cast Al-Si alloys: A review. *Mater. Today Proc.* **2022**, *62*, 6568–6572. [CrossRef]
6. Hart, G.L.W.; Mueller, T.; Toher, C.; Curtarolo, S. Machine learning for alloys. *Nat. Rev. Mater.* **2021**, *6*, 730–755. [CrossRef]
7. Fu, K.; Zhu, D.; Zhang, Y.; Zhang, C.; Wang, X.; Wang, C.; Jiang, T.; Mao, F.; Zhang, C.; Meng, X.; et al. Predictive Modeling of Tensile Strength in Aluminum Alloys via Machine Learning. *Materials* **2023**, *16*, 7236. [CrossRef] [PubMed]
8. Hao, C.; Sui, Y.; Yuan, Y.; Li, P.; Jin, H.; Jiang, A. Composition optimization design and high temperature mechanical properties of cast heat-resistant aluminum alloy via machine learning. *Mater. Des.* **2025**, *250*, 113587. [CrossRef]
9. Murakami, Y.; Furushima, R.; Shiga, K.; Miyajima, T.; Omura, N. Mechanical property prediction of aluminium alloys with varied silicon content using deep learning. *Acta Mater.* **2025**, *286*, 120683. [CrossRef]
10. Zhao, Y.; He, Z.; Wang, D.; Liu, X.; Chao, Y.; Ye, J.; Zhong, Y.; Lu, Z.; Cheng, L.; Hu, C. Machine learning-guided design of aluminum–silicon alloys with balanced strength and ductility. *J. Mater. Res. Technol.* **2026**, *43*, 3098–3109. [CrossRef]

11. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. [CrossRef] [PubMed]
12. Li, K.; Rubungo, A.N.; Lei, X.; Persaud, D.; Choudhary, K.; DeCost, B.; Dieng, A.B.; Hattrick-Simpers, J. Probing out-of-distribution generalization in machine learning for materials. *Commun. Mater.* **2025**, *6*, 9. [CrossRef]
13. Witman, M.D.; Schindler, P. MatFold: Systematic insights into materials discovery models' performance through standardized cross-validation protocols. *Digit. Discov.* **2025**, *4*, 625–635. [CrossRef]
14. Hu, J.; Liu, D.; Fu, N.; Dong, R. Realistic material property prediction using domain adaptation based machine learning. *Digit. Discov.* **2024**, *3*, 300–312. [CrossRef]
15. Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R.J.; Wasserman, L. Distribution-Free Predictive Inference for Regression. *J. Am. Stat. Assoc.* **2018**, *113*, 1094–1111. [CrossRef]
16. Angelopoulos, A.N.; Bates, S. Conformal Prediction: A Gentle Introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [CrossRef]
17. Yang, Y.; Kuchibhotla, A.K.; Tchetgen Tchetgen, E. Doubly robust calibration of prediction sets under covariate shift. *J. R. Stat. Soc. Ser. B* **2024**, *86*, 943–965. [CrossRef] [PubMed]
18. Geifman, Y.; El-Yaniv, R. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. *Proc. Mach. Learn. Res.* **2019**, *97*, 2151–2159.
19. Jarfors, A.E.W.; Jansson, P. Selecting Cast Alloy Alloying Elements Suitable for a Circular Society. *Sustainability* **2022**, *14*, 6584. [CrossRef]
20. Bogdanoff, T.; Tiryakioğlu, M.; Liljenfors, T.; Jarfors, A.E.W.; Seifeddine, S.; Ghassemali, E. On the Effectiveness of Rotary Degassing of Recycled Al-Si Alloy Melts: The Effect on Melt Quality and Energy Consumption for Melt Preparation. *Sustainability* **2023**, *15*, 5189. [CrossRef]
21. Scharf, S.; Bergedieck, N.; Riedel, E.; Richter, H.; Stein, N. Unlocking Sustainability Potentials in Heat Treatment Processes. *Sustainability* **2020**, *12*, 6457. [CrossRef]
22. Cai, J.; Han, M.; Yan, X.; Chen, Y.; Li, D.; Zhao, K.; Zhang, D.; Hu, K.; Sua, H.H.; Jun, H.K.; et al. A process-synergistic active learning framework for high-strength Al–Si alloys design. *npj Comput. Mater.* **2025**, *11*, 228. [CrossRef]
23. Bergmann Kremer, A.; Scheidhauer, N.; Szucki, M.; Keßler, A.; Krawczyk, L.; Dietrich, I. Influence of the Chemical Composition, Casting Process, Grain Refinement, and Modification on the Mechanical and Corrosion Behaviours of Al-Si Casting Alloys. *Int. J. Met.* **2026**. [CrossRef]
24. Hillebrandt-Szymańska, K.K.; Piątkowski, J. The Influence of Overheating Temperature on the Shape Change of Primary Silicon Crystals and the Mechanical Properties of AlSi17 Alloy. *Arch. Foundry Eng.* **2025**, *25*, 94–101. [CrossRef]
25. Stodden, V.; Seiler, J.; Ma, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 2584–2589. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.