

(RESEARCH ARTICLE)



GRADE-AWARE MULTI-TARGET MACHINE LEARNING OF STEEL MECHANICAL PROPERTIES: ENSEMBLE, ANN, AND SELF-ATTENTION TRANSFORMER MODELS UNDER UNSEEN-GRADE SHIFT

Harikrishnan Kumarasamy *

Independent Researcher, Greek row Dr, Arlington, Texas, USA.

* Corresponding Author; Email: hari_gce@yahoo.co.in

ORCID Details

Harikrishnan Kumarasamy: <https://orcid.org/0000-0002-6010-849X>

Global Journal of Engineering and Technology Advances, , 2026, 29(01), 319–331

Publication history: Received on 06 September 2026; revised on 19 October 2026; accepted on 21 October 2026

Article DOI: <https://doi.org/10.30574/gjeta.2026.29.1.0300>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Predicting steel mechanical properties from heterogeneous chemistry and heat-treatment metadata is a domain-transfer problem because repeated observations are hierarchically clustered by grade and property reporting is incomplete. This study develops a leakage-resistant multi-target learning framework using 1,360 Steel Bench records representing 562 unique grades, 12 chemistry and processing descriptors, and five responses: tensile strength, yield strength, elongation, reduction of area, and impact energy. Missing targets are not imputed. Numeric imputation, standardization, and categorical encoding g are fitted within each training fold, while neural models use a masked multi-response objective. Twelve conventional regressors are benchmarked against a 256-128-64 artificial neural network and a two-block, four-head self-attention Transformer. Model selection uses three repeats of five-fold grade-grouped cross-validation, followed by a completely grade-disjoint locked test. Extra Trees provides strong cross-target development performance, with grouped-cross-validation R^2 values of approximately 0.572, 0.462, 0.558, 0.366, and 0.354 for tensile strength, yield strength, elongation, reduction of area, and impact energy, respectively, while Gradient Boosting leads the development ranking for elongation and impact energy. On the locked unseen-grade test, elongation remains transferable ($R^2 = 0.646$) and impact energy retains moderate skill ($R^2 = 0.370$), whereas tensile strength, yield strength, and reduction of area exhibit negative R^2 . Distribution-shift and descriptor-ablation analyses show that chemistry redistribution, missingness shift, and incomplete process-state descriptors constrain transfer more strongly than nominal model complexity.

Keywords: Steel; Mechanical Properties; Materials Informatics; Grade Shift; Grouped Cross-Validation; Self-Attention

1 INTRODUCTION

Artificial intelligence and machine learning (AI/ML) are increasingly important across diverse alloy and functional-material systems because complex relationships among composition, processing, structure, and properties are difficult to characterize exhaustively through experiments alone. Data-driven materials models provide a systematic approach for learning these multidimensional relationships and predicting material properties from available composition and

processing descriptors. Materials research encompasses systems with diverse crystallographic and phase structures, including body-centered cubic (BCC), face-centered cubic (FCC), hexagonal close-packed (HCP), and ordered structures, further increasing the complexity of structure–property relationships. Experimental investigations across alloy systems ([1]–[4]) and functional materials [5] demonstrate the sensitivity of material response to composition, processing conditions, phase constitution, and material state. AI/ML provides an additional computational framework for extracting relationships from such multidimensional materials data. In related work, AI/ML-assisted inverse alloy design using tree-ensemble and deep neural-network surrogate models was applied to multi-principal-element alloys for targeted high-temperature properties [6]. In contrast, the present study addresses the forward prediction problem of estimating steel mechanical properties from chemistry and heat-treatment descriptors, with particular emphasis on model reliability and generalization to previously unseen steel grades.

For high-dimensional alloy systems, data-driven surrogates can learn nonlinear trends, screen candidate compositions, and prioritize experiments while avoiding the cost of exhaustive trial-and-error testing. General materials-informatics methods provide the computational foundation [6], [7], [8]. Kumarasamy’s AI/ML-assisted inverse alloy-design study of multi-principal-element alloys used tree ensembles and deep neural-network surrogates to screen candidate compositions against targeted high-temperature mechanical-property requirements [9]. That study addressed inverse alloy selection, whereas the present work investigates whether steel mechanical-property models remain reliable when transferred to previously unseen grades. Both problems require leakage-resistant validation, sufficient chemistry and processing descriptors, and explicit recognition of the limitations of predictions made outside well-supported data domains. Accordingly, the present study examines missing property labels, grade-grouped validation, and an untouched grade-disjoint test population.

Steel is an especially demanding test case for this requirement. A tabular chemistry and heat-treatment model is necessarily a projection of a higher-dimensional metallurgical state onto an incomplete descriptor manifold. The available descriptors may omit cooling trajectory, hold time, prior deformation, phase fractions, grain size, specimen geometry, and other variables that determine mechanical response. Consequently, increasing algorithmic complexity alone cannot guarantee transfer when the evaluation population differs chemically or metallurgically from the development population.

A second difficulty is hierarchical dependence. Multiple observations associated with the same nominal grade are not statistically independent. Row-randomized train/test partitioning can therefore place chemically near-identical grade instances on both sides of the split, converting a nominal generalization experiment into interpolation. The present work treats grade identity as a non-predictive grouping variable and asks a more stringent question: can a model trained on one set of grades transfer to grades that are absent from model development?

The contribution is a unified benchmark rather than a single neural architecture. Twelve conventional regressors, a shared-representation multi-task ANN, and a genuine feature-token Transformer are evaluated under identical data governance. Missing response values remain missing, model selection is confined to grouped development data, and the final test set is grade-disjoint. The resulting performance gap between repeated grouped cross-validation and locked testing is treated as a scientific result about descriptor sufficiency and distribution shift, not as an error to be hidden.

2 MATERIALS AND METHODS

2.1 Dataset, descriptors, and response tensor

The present study uses SteelBench v1.0[10], a publicly available dataset for data-driven prediction of steel mechanical properties. The analyzed cohort contains 1,360 records and 562 unique grade identifiers. Twelve predictors are retained: C, Mn, Cr, Mo, Ni, Si, V, Cu, Al, austenitizing temperature, quench medium, and tempering temperature. The response vector contains tensile strength, yield strength, elongation, reduction of area, and average impact energy. Hardness is excluded because the source field mixes heterogeneous hardness scales and strings that require a separate harmonization study. Figure 1 illustrates the observation topology and missing-property pattern of the analyzed chemistry–process–property data, while Figure 2 shows the grade-disjoint training, validation, and locked-test populations.

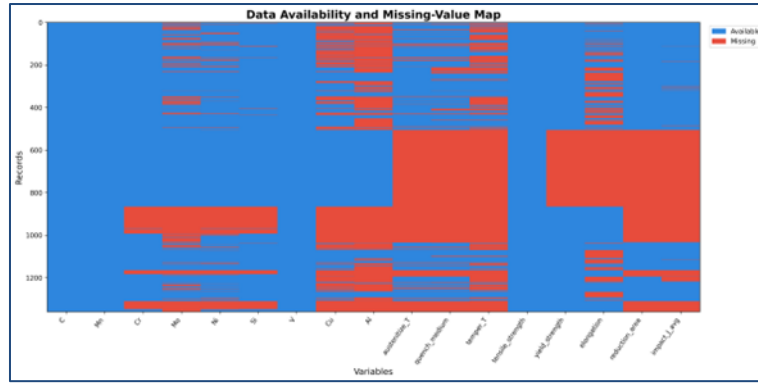


Figure 1: Observation topology of the partially observed chemistry–process–property tensor. Missing targets are retained rather than synthetically reconstructed

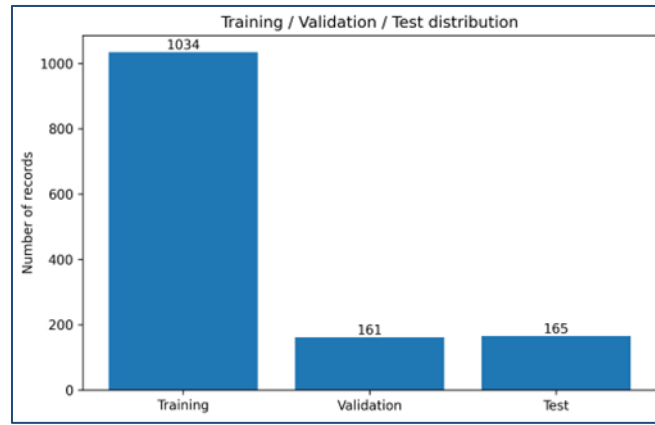


Figure 2: Grade-disjoint training, validation, and locked-test populations. Grade identity is used only for grouping.

2.2 Fold-local preprocessing and incomplete-label learning

Numeric missing inputs are replaced by the training-fold median and standardized using statistics learned exclusively from the training fold. Quench medium is mode-imputed and one-hot encoded. For feature j following standard feature scaling [11]:

$$z_j = \frac{x_j - \mu_j}{\sigma_j} \quad (1)$$

Neural targets are standardized only where observed. A binary availability tensor m_{ik} gates the contribution of target k for observation i . The study-specific masked objective is

$$\mathcal{L}_{\text{mask}} = \frac{\sum_i \sum_k m_{ik} (\hat{y}_{ik} - y_{ik})^2}{\sum_i \sum_k m_{ik}} \quad (2)$$

Thus, unreported targets generate neither pseudo-labels nor gradients. This is essential because property availability is strongly nonuniform across the source data.

2.3 Computational process flow

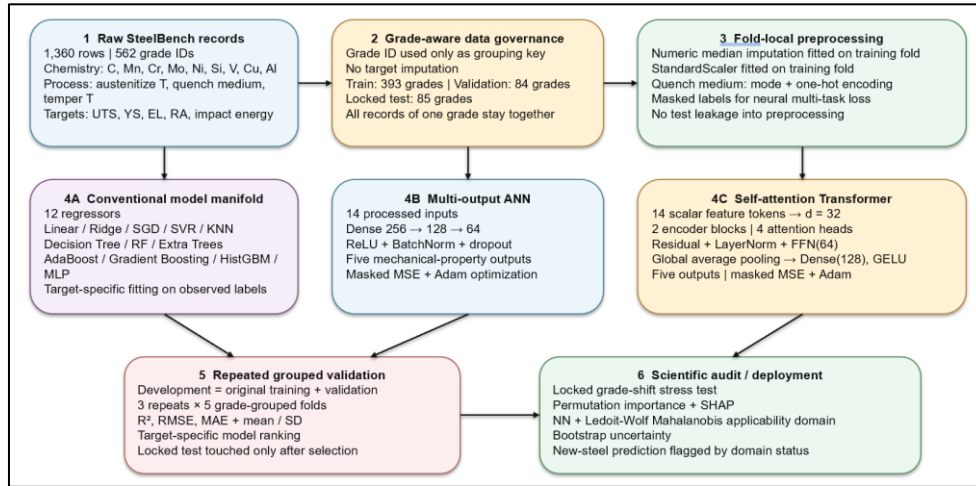


Figure 3: End-to-end computational workflow for “Leakage-resistant multi-target steel-property learning workflow”: data governance, fold-local preprocessing, conventional/ANN/Transformer model manifolds, repeated grade-grouped selection, and locked-test release.

2.4 Predictive model manifold

2.4.1 Conventional nonlinear regressors

The benchmark includes linear regression, RidgeCV, stochastic-gradient regression, support-vector regression, k-nearest neighbors, decision tree, random forest, Extra Trees, AdaBoost, gradient boosting, histogram gradient boosting, and a scikit-learn multilayer perceptron. Random Forest [12], Extra Trees [13], and Gradient Boosting [14] are of particular interest because they represent different bias–variance and randomization strategies for nonlinear tabular regression. Figure 3 illustrates the complete computational workflow, including grade-aware data governance, fold-local preprocessing, model development, repeated grouped validation, and locked-test evaluation.

2.4.2 Multi-task artificial neural network

$$\mathbf{h}^{(\ell)} = \phi(\mathbf{W}^{(\ell)}\mathbf{h}^{(\ell-1)} + \mathbf{b}^{(\ell)}) \quad (3)$$

$$\hat{x} = \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}, \quad y = \gamma\hat{x} + \beta \quad (4)$$

Optimization uses Adam [16], with dropout regularization following Srivastava et al. [17].

2.4.3 Feature-token Transformer

The Transformer reshapes the processed vector into a 14-token scalar sequence, projects each token to a 32-dimensional latent representation, and applies two encoder-style blocks containing four-head self-attention, residual pathways, layer normalization, and 64-unit position-wise feed-forward transformations. Global average pooling and a 128-unit GELU layer feed the five-property output. The model contains 22,021 trainable parameters. The attention mechanism follows the Transformer formulation of Vaswani et al. [18].

$$\mathbf{Q} = \mathbf{XW}_Q, \quad \mathbf{K} = \mathbf{XW}_K, \quad \mathbf{V} = \mathbf{XW}_V \quad (5)$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (6)$$

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)\mathbf{W}_O \quad (7)$$

2.5 Grade-grouped model selection

The deterministic initial segregation contains 1,034 training rows from 393 grades, 161 validation rows from 84 grades, and 165 locked-test rows from 85 grades. Reviewer-level selection combines training and validation into a development

cohort and performs three independently randomized repetitions of five-fold grade-grouped cross-validation. The locked test is not used for model ranking. For metric q over K folds/repeats,

$$\bar{q} = \frac{1}{K} \sum_{k=1}^K q_k \quad (8)$$

$$s_q = \sqrt{\frac{1}{K-1} \sum_{k=1}^K (q_k - \bar{q})^2} \quad (9)$$

Target-specific ranks are computed independently for descending R^2 and ascending RMSE and MAE, then summarized by the study-specific composite

$$R_{\text{comp}} = \frac{R_{R^2} + R_{\text{RMSE}} + R_{\text{MAE}}}{3} \quad (10)$$

3 RESULTS AND DISCUSSION

3.1 Descriptor-response topology

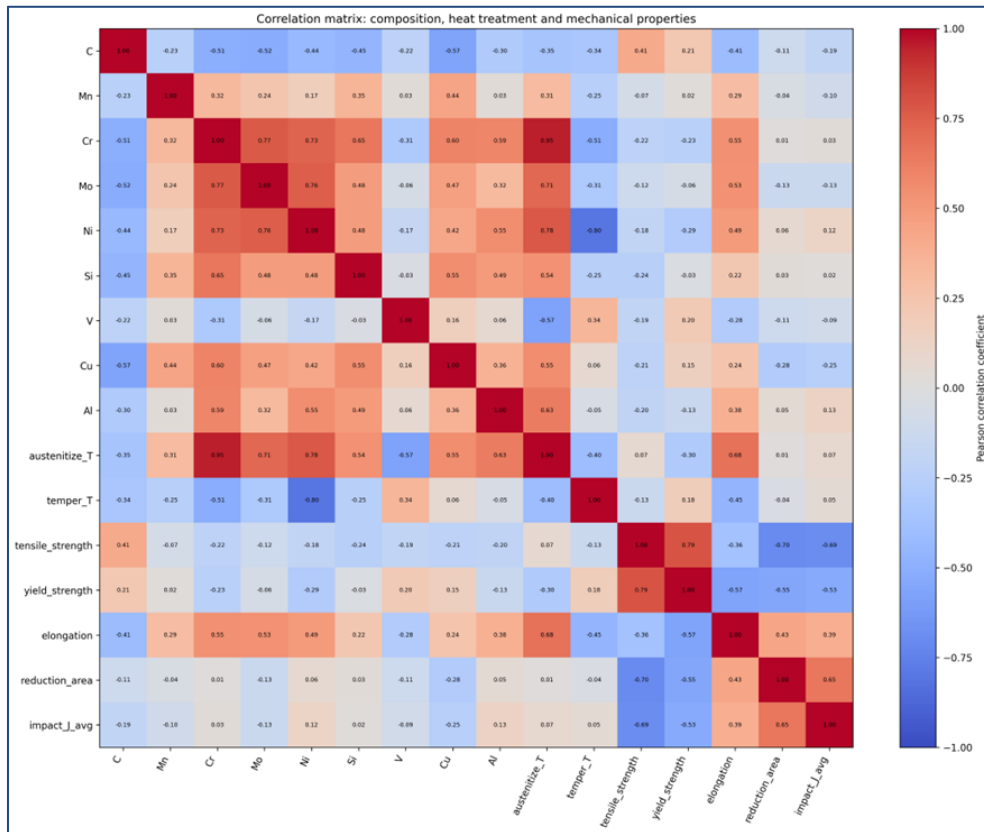


Figure 4: Correlation topology across chemistry, processing variables, and five mechanical responses.

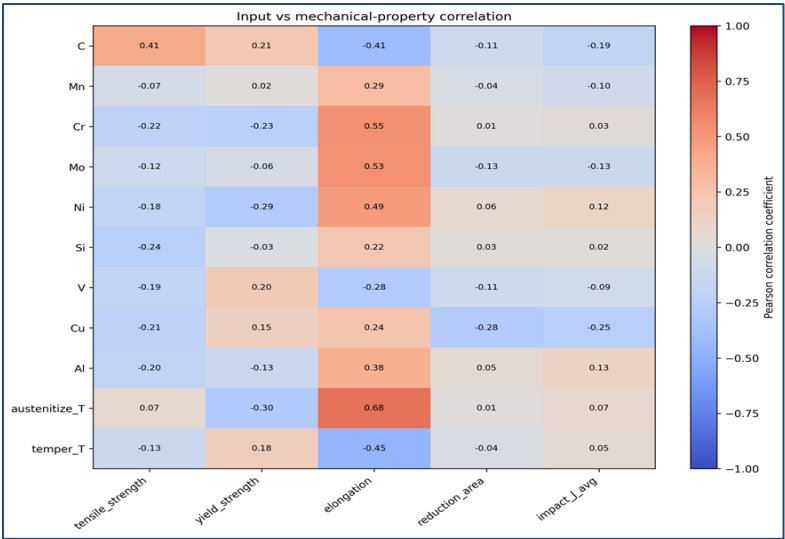


Figure 5: Input-to-target correlation subspace. Pairwise sample sizes vary because missing observations are preserved.

The strongest first-order associations include C with tensile strength ($r \approx 0.407$), austenitizing temperature with elongation ($r \approx 0.679$), and Cu with both reduction of area and impact energy in the available paired observations. These coefficients are descriptive and cannot be interpreted causally because alloy-family, heat-treatment, and source covariates are entangled. Figures 4 and 5 illustrate the overall descriptor–response correlation topology and the corresponding input-to-target correlation subspace, respectively.

3.2 Repeated grouped-CV model landscape

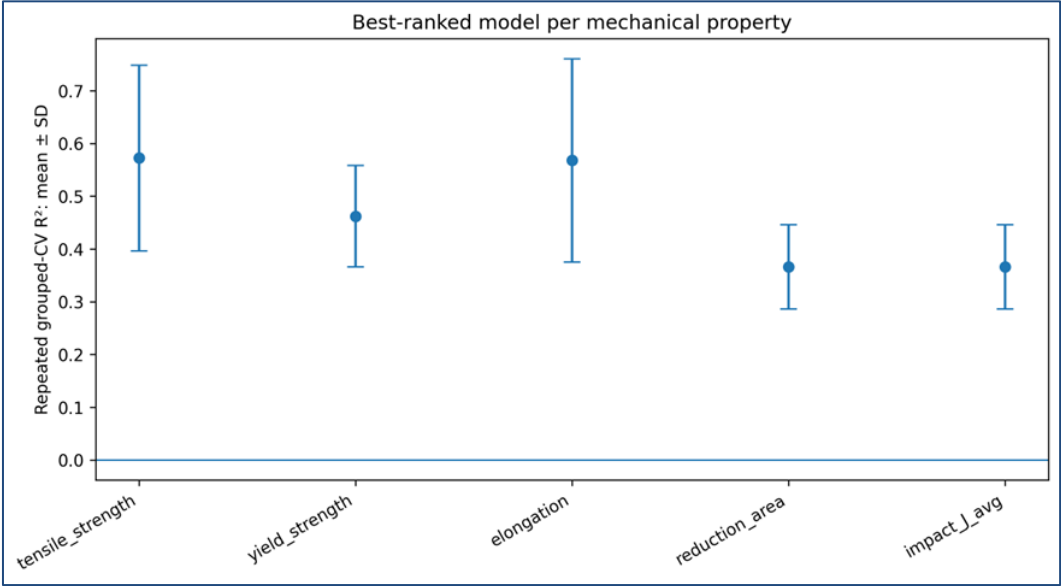


Figure 6: Mean grouped-CV R² and fold-to-fold dispersion for the target-specific best-ranked models.

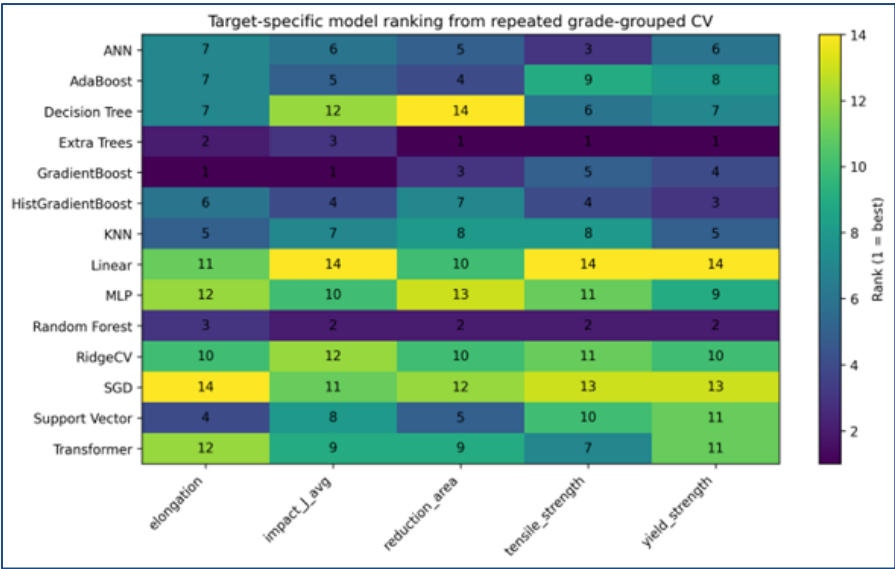


Figure 7: Unified target-specific model-ranking landscape across conventional regressors, ANN, and Transformer.

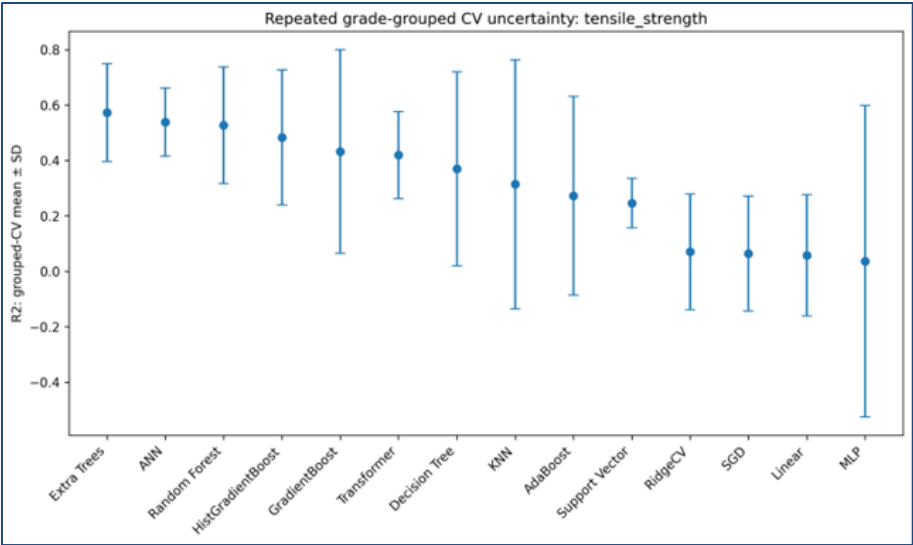


Figure 8: Tensile-strength grouped-CV R² distribution, illustrating partition-induced epistemic variability.

Extra Trees is the highest-ranked model for tensile strength, yield strength, and reduction of area. Gradient Boosting leads elongation and impact-energy ranking. Extra Trees achieves mean grouped-CV $R^2 \approx 0.572$ for tensile strength and ≈ 0.462 for yield strength. Elongation is the most consistently transferable response during development, with Gradient Boosting reaching $R^2 \approx 0.568$ and Extra Trees ≈ 0.558 . The spread across grouped folds is non-negligible, demonstrating that the identity of the withheld grades materially perturbs the apparent model quality. Figures 6–8 summarize the repeated grouped-cross-validation results, including target-specific performance and dispersion, model rankings, and tensile-strength variability across grouped partitions.

3.3 Neural optimization and architecture comparison

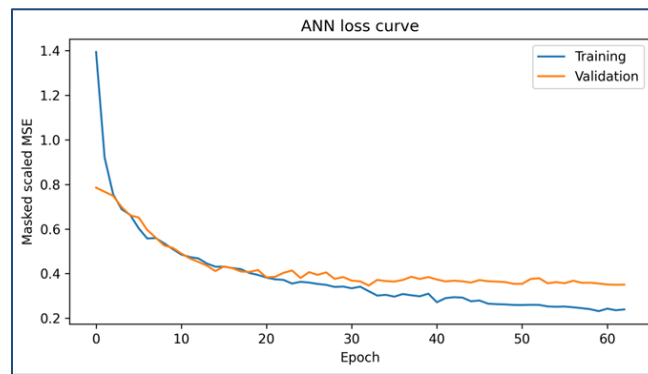


Figure 9: Representative masked multi-output ANN optimization trajectory.

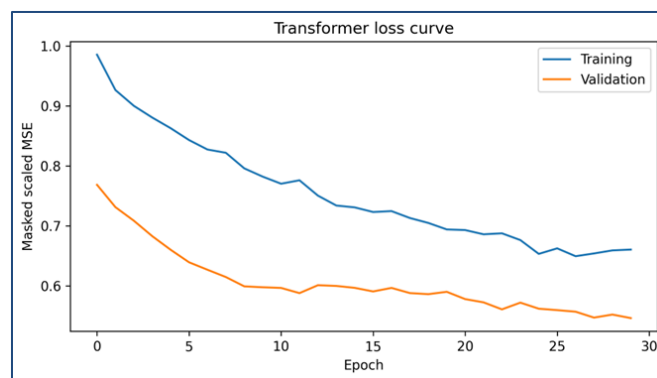


Figure 10: Representative Transformer optimization trajectory.

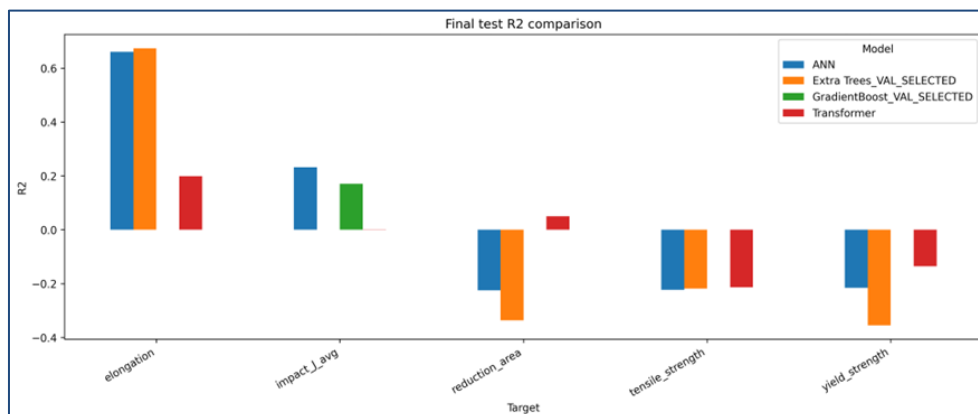


Figure 11: Primary test-set R^2 comparison among conventional, ANN, and Transformer model families.

The ANN is competitive for tensile strength but does not dominate the ensemble regressors across responses. The Transformer is a genuine self-attention model, yet its additional representational flexibility does not translate into systematic advantage for the moderate-size sparse tabular cohort. This behavior is consistent with a regime in which grade-level distribution shift and incomplete metallurgical state descriptors dominate the error budget more strongly than architectural capacity. Figures 9 and 10 show the representative optimization trajectories of the ANN and Transformer, respectively, while Figure 11 compares their locked-test R^2 performance with the selected conventional models.

3.4 Locked unseen-grade stress test

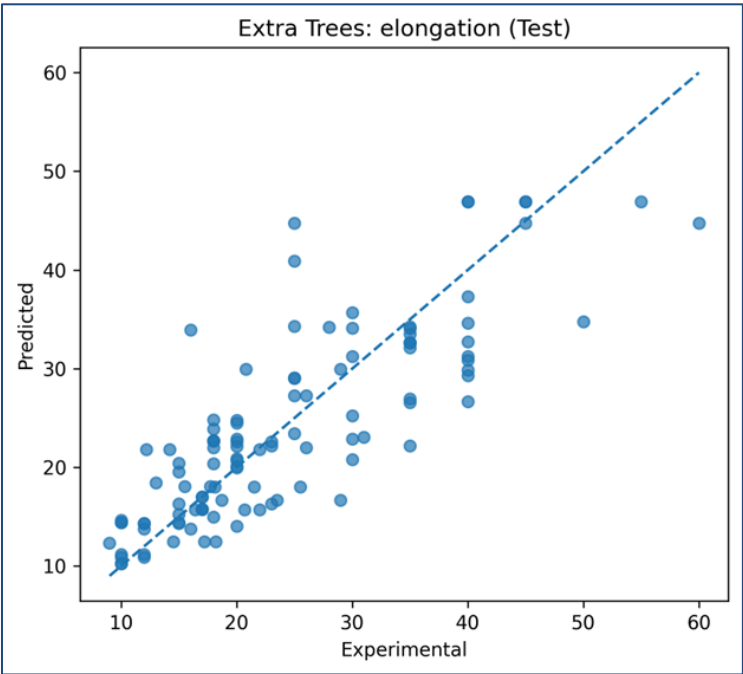


Figure 12: Locked-test actual versus predicted elongation for Extra Trees.

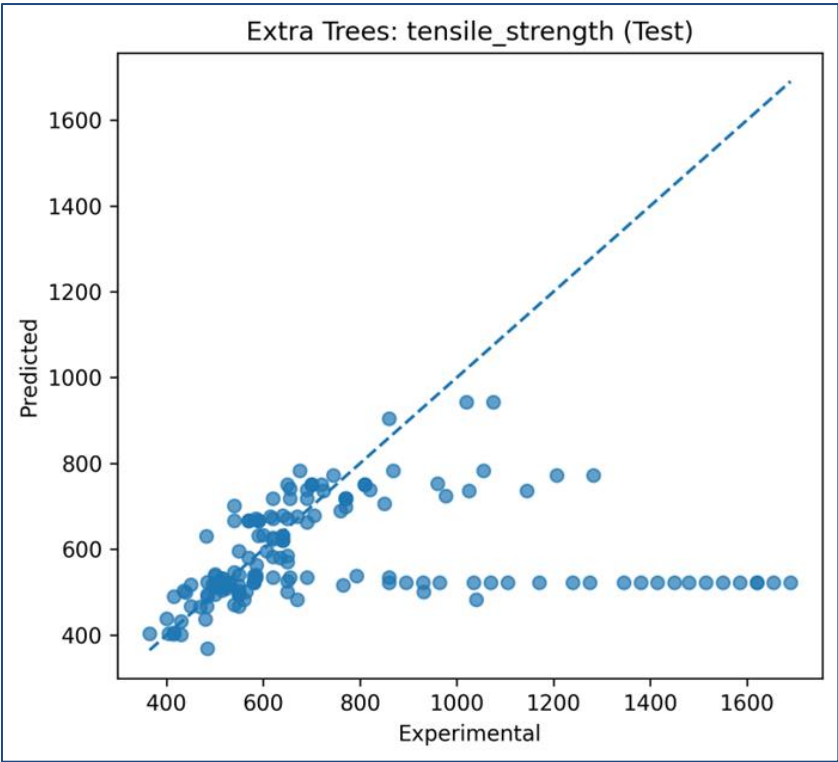


Figure 13: Locked-test tensile-strength actual versus predicted response, showing severe grade-shift degradation.

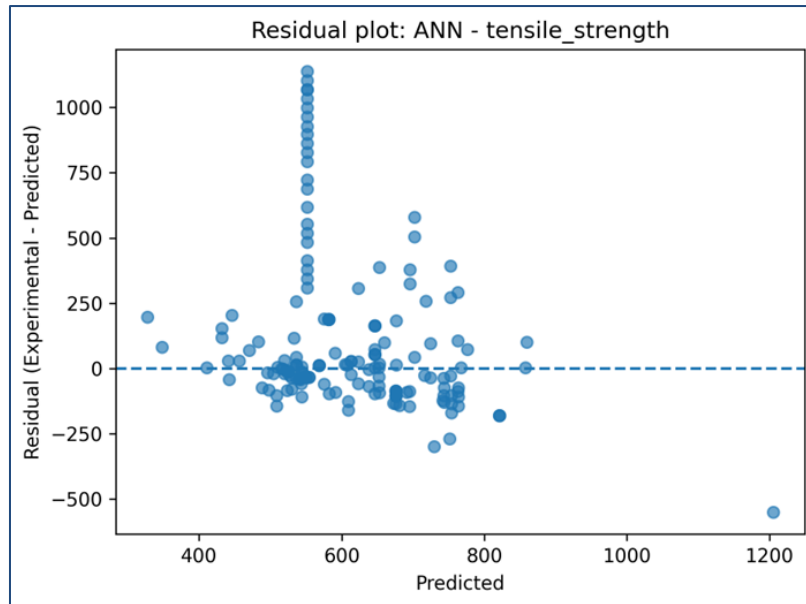


Figure 14: ANN tensile-strength residual structure on the primary test partition.

After selection exclusively on development data, elongation remains strongly transferable on the locked test ($R^2 = 0.646$; $RMSE \approx 6.42$; $MAE \approx 4.79$), while impact energy retains moderate skill ($R^2 = 0.370$; $RMSE \approx 28.99$ J; $MAE \approx 23.88$ J). In contrast, tensile strength ($R^2 = -0.224$), yield strength ($R^2 = -0.376$), and reduction of area ($R^2 = -0.284$) fall below the constant-mean baseline. These negative R^2 values are retained because they diagnose a genuine generalization failure rather than a plotting artifact. Figures 12 and 13 compare locked-test actual and predicted responses for elongation and tensile strength, respectively, while Figure 14 shows the ANN tensile-strength residual structure.

3.5 General discussion

The dominant result is the divergence between repeated grade-grouped development performance and the locked unseen-grade stress test. The difference establishes that favorable cross-validation statistics can coexist with severe transfer failure when the evaluation population occupies a shifted region of the alloy/process manifold. A random row-wise split would answer a less demanding interpolation question and can conceal this failure mode.

The strong performance of Extra Trees reflects a favorable inductive bias for heterogeneous nonlinear tabular data: randomized split thresholds capture interaction structure while ensemble averaging suppresses variance. By contrast, self-attention must estimate a substantially more structured latent interaction operator from a relatively small and incompletely labeled cohort. The result should not be read as a general rejection of Transformers in metallurgy; rather, it indicates that attention is not a substitute for missing phase, microstructure, thermal-history, and test-condition information.

The response dependence is also important. Elongation and impact energy retain meaningful locked-test skill, whereas strength transfer deteriorates strongly. This suggests that the sufficiency of the current descriptor set is target dependent. Future models should incorporate microstructural state variables, cooling rate, tempering time, prior deformation, specimen geometry, phase fractions, and source metadata before increasing architecture complexity.

3.6 Quantifying the locked-test distribution shift

The deterioration from repeated grade-grouped development performance to the locked unseen-grade test was examined directly rather than attributed generically to “distribution shift.” Two-sample Kolmogorov–Smirnov analysis demonstrates statistically significant marginal redistribution for most alloy descriptors. The largest shifts occur for Al ($D = 0.662$, $p < 10^{-13}$), Cu ($D = 0.484$, $p < 10^{-10}$), Mo ($D = 0.407$, $p < 10^{-12}$), Ni ($D = 0.366$, $p < 10^{-12}$), Si ($D = 0.346$, $p < 10^{-11}$), and Cr ($D = 0.337$, $p < 10^{-11}$). Standardized mean differences are particularly large for Cr (0.670), Mo (0.652), Ni (0.647), and Si (0.598). The locked set is therefore not merely grade-disjoint; its chemistry distribution is materially displaced from the development population.

Missingness also shifts. For example, Cu missingness rises from 38.8% in development to 69.1% in the locked test, Al from 54.5% to 80.0%, and Mo from 27.3% to 42.4%. Because median imputation is learned only from development data, this change increases dependence on training-distribution summary statistics precisely in descriptors whose

observed distributions have also shifted. This combination provides a concrete mechanism for degradation under the locked test.

The population composition changes at the metadata level as well. Source, data tier, steel family, and quench medium all differ significantly between development and locked-test populations ($\chi^2 p < 3 \times 10^{-6}$ in every case). Cramér's V is 0.229 for source, 0.229 for data tier, 0.258 for steel family, and 0.145 for quench medium. Most notably, NIMS records constitute 30.1% of the development cohort but are absent from the locked test, whereas EMK_{spec} increases from 53.3% to 70.3%. Austenitic stainless steels increase from 11.4% to 30.9% of the locked population. These shifts provide quantitative evidence that the final test represents a different mixture of alloy families and provenance rather than a random sample from the development distribution. Figure 15 illustrates the magnitude of the development-to-locked-test distribution shift across the continuous chemistry and processing descriptors.

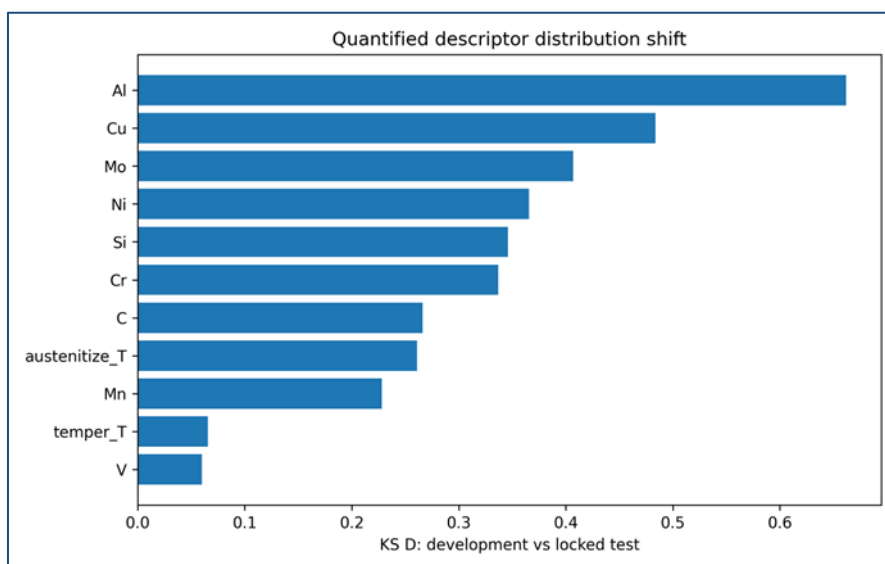


Figure 15: Quantified development-to-locked-test descriptor shift. Bars report the two-sample Kolmogorov-Smirnov D statistic for each continuous chemistry/process descriptor; larger values indicate stronger marginal distributional displacement.

3.7 Descriptor ablation: chemistry versus processing information

A targeted ablation was performed with Extra Trees under the same three-repeat, five-fold grade-grouped validation logic to determine whether predictive information is primarily carried by alloy chemistry or by the available heat-treatment descriptors. Three representations were evaluated: chemistry only (C, Mn, Cr, Mo, Ni, Si, V, Cu, Al), processing only (austenitizing temperature, tempering temperature, quench medium), and chemistry plus processing.

Chemistry alone retains most of the predictive information. Its mean grouped-CV R^2 values are 0.530 for tensile strength, 0.428 for yield strength, 0.622 for elongation, 0.311 for reduction of area, and 0.300 for impact energy. Processing-only performance is substantially lower: 0.188, 0.121, 0.198, 0.089, and 0.113, respectively. Adding the available processing variables to chemistry produces only modest and target-dependent changes: tensile $R^2 = 0.517$, yield = 0.440, elongation = 0.659, reduction of area = 0.300, and impact energy = 0.301. The corresponding numerical results are summarized in Table 1.

The ablation clarifies the source of the locked-test vulnerability. The current heat-treatment fields do not supply enough independent information to stabilize prediction when alloy-family chemistry shifts. For elongation, processing information adds measurable value to chemistry, increasing mean R^2 from 0.622 to 0.659. For tensile strength and reduction of area, however, the combined representation is not better than chemistry alone within the observed fold variability. Thus, simply increasing model complexity cannot compensate for incomplete process-state descriptors; richer variables such as hold time, cooling trajectory, prior deformation, phase fraction, grain size, and tempering history are more likely to improve transferability. Figure 16 summarizes the chemistry-only, processing-only, and combined descriptor-ablation results under repeated grade-grouped validation.

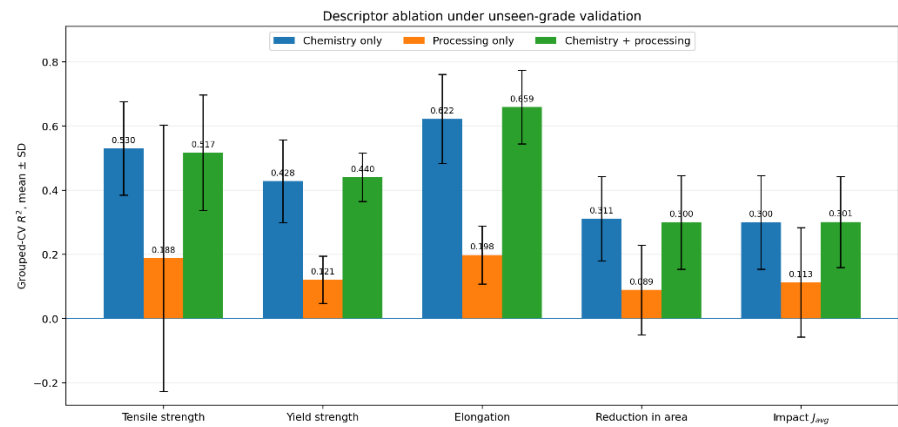


Figure 16: Descriptor ablation under repeated grade-grouped validation using Extra Trees. Bars show mean $R^2 \pm$ standard deviation across 15 grouped runs for chemistry-only, processing-only, and combined representations.

Table 1: Grade-grouped descriptor ablation. Values are mean $R^2 \pm$ SD across 15 runs.

Target	Chemistry (R^2)	Processing (R^2)	Combined (R^2)
Tensile strength	0.530 ± 0.146	0.188 ± 0.415	0.517 ± 0.180
Yield strength	0.428 ± 0.129	0.121 ± 0.074	0.440 ± 0.075
Elongation	0.622 ± 0.139	0.198 ± 0.090	0.659 ± 0.115
Reduction in area	0.311 ± 0.132	0.089 ± 0.140	0.300 ± 0.146
Impact J_{avg}	0.300 ± 0.146	0.113 ± 0.171	0.301 ± 0.142

4 CONCLUSION

A leakage-resistant five-response steel-property benchmark was established using grade-disjoint data governance, incomplete-label preservation, twelve conventional regressors, a multi-task ANN, and a genuine self-attention Transformer. Repeated 3×5 grade-grouped validation identifies tree ensembles as the most robust model family for the present tabular representation. The locked test, however, reveals a decisive response-dependent generalization penalty: elongation and impact energy retain positive transfer, while tensile strength, yield strength, and reduction of area fail to outperform a mean baseline. The study therefore reframes steel-property ML from a model-capacity problem to a descriptor-sufficiency and domain-transfer problem. The protocol provides a reproducible baseline for future physics-enriched and multimodal steel informatics.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Disclosure of conflict of interest

The author declare no conflicts of interest regarding this manuscript.

REFERENCES

[1] K. Harikrishnan, K. Chandra, P. S. Misra, V. S. Agarwala, In-situ XRD investigations of two step R-phase transformations in aged NiTi shape memory alloys, ESOMAT 2009, 02018 (2009). DOI: 10.1051/esomat/200902018.

[2] K. Harikrishnan, K. Chandra, P. S. Misra, V. S. Agarwala, B19 orthorhombic martensitic transformations in aged TiNiCu shape memory alloys, ESOMAT 2009, 02020 (2009). DOI: 10.1051/esomat/200902020.

[3] K. Hari Krishnan, P. S. Misra, K. Chandra, V. S. Agarwala, Low Temperature Ageing Effects in NiTi Shape Memory Alloys, International Journal of Material Science 4(2) (2014) 73-77. DOI: 10.14355/ijmsci.2014.0402.05.

- [4] K. Harikrishnan, P. S. Misra, K. Chandra, V. S. Agarwala, Characteristic of Burst Transformations in Pseudoelasticity and Shape Memory Effect - a Review, ESOMAT 2009, 02019 (2009). DOI: 10.1051/esomat/200902019.
- [5] K. Harikrishnan, D. V. Bavbande, D. Mohan, B. Manoharan, M. R. S. Prasad, G. Kalyanakrishnan, PZT/PLZT-elastomer composites with improved piezoelectric voltage coefficient, IOP Conference Series: Materials Science and Engineering 314(1) (2018) 012007. DOI: 10.1088/1757-899X/314/1/012007.
- [6] L. Ward et al., A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput. Mater. 2 (2016) 16028. DOI: 10.1038/npjcompumats.2016.28.
- [7] R. Ramprasad et al., Machine learning in materials informatics: recent applications and prospects, npj Comput. Mater. 3 (2017) 54. DOI: 10.1038/s41524-017-0056-5.
- [8] A. Agrawal, A. Choudhary, Materials informatics and big data, APL Mater. 4 (2016) 053208. DOI: 10.1063/1.4946894.
- [9] H. Kumarasamy, AI/ML-Assisted Inverse Alloy Design of Multi-Principal Element Alloys for High-Temperature Energy Systems: Comparative Tree Ensembles and Deep Neural-Network Surrogates, Journal of Energy Research and Reviews 18(10) (2026) 67–80. DOI: 10.9734/jenrr/2026/v18i10543.
- [10] A. Volkov et al., SteelBench v1.0: A Benchmark Dataset for Steel Mechanical Property Prediction with Grade-Shift Evaluation, Zenodo (2026). DOI: 10.5281/zenodo.18530558.
- [11] F. Pedregosa et al., Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
- [12] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32. DOI: 10.1023/A:1010933404324.
- [13] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine Learning 63 (2006) 3–42. DOI: 10.1007/s10994-006-6226-1.
- [14] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Stat. 29 (2001) 1189–1232. DOI: 10.1214/aos/1013203451.
- [15] S. Ioffe, C. Szegedy, Batch Normalization, Proc. ICML (2015) 448–456.
- [16] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, ICLR (2015), arXiv:1412.6980.
- [17] N. Srivastava et al., Dropout: A Simple Way to Prevent Neural Networks from Overfitting, JMLR 15 (2014) 1929–1958.
- [18] A. Vaswani et al., Attention Is All You Need, NeurIPS 30 (2017), arXiv:1706.03762.