Nonlinear Machine Learning for Compositional Superconductivity Prediction

David Szczecina
University of Waterloo
david.szczecina@uwaterloo.ca

Tanjeem Hossain
University of Waterloo
tanjeem.hossain@uwaterloo.ca

Kamal Ahsan
University of Waterloo
k4ahsan@uwaterloo.ca

***Abstract*—Accurate prediction of superconducting critical temperature (T_c) from chemical composition is a central challenge in materials informatics and an important step toward accelerating superconductor discovery. In this work, we conduct a systematic benchmarking study of linear regression, regularized linear models, tree-based ensembles, gradient boosting methods, neural networks, and stacked ensembles on the UCI Superconductivity dataset comprising over 21,000 materials described by engineered compositional descriptors derived from elemental properties.**

Our results demonstrate that nonlinear models substantially outperform linear baselines, indicating that T_c depends on complex, non-additive interactions among compositional descriptors. Tree-based ensemble methods consistently achieve the strongest predictive performance, with random forest models achieving the lowest validation error (RMSE $\approx$ 9 K). Boosting approaches and stacked ensembles provide competitive alternatives, while neural networks capture nonlinear structure but do not outperform tree ensembles in this setting. Cross-validation analysis reveals increased variability in flexible nonlinear models relative to linear regression, highlighting the bias–variance tradeoff inherent in high-capacity learners. Overall, these findings emphasize the importance of nonlinear modeling for compositional materials prediction and demonstrate that ensemble learning methods provide robust and effective predictive performance for superconductivity discovery tasks.

I. Introduction

The discovery of superconducting materials with elevated critical temperatures remains one of the most enduring challenges in condensed matter physics [1]. Since the identification of high-temperature cuprate superconductors, researchers have sought to understand the mechanisms governing superconductivity and to identify new materials capable of sustaining superconducting states at increasingly practical temperatures. Superconductors enable transformative technologies including lossless power transmission, high-field electromagnets, particle accelerators, and quantum information systems. However, the experimental search for novel superconductors is inherently expensive and time-intensive, requiring complex synthesis, structural characterization, and cryogenic measurements.

Computational methods offer a promising pathway to accelerate this discovery process. In particular, data-driven machine learning approaches have demonstrated the ability to approximate complex structure–property relationships directly from curated datasets [2]. Predicting superconducting critical temperature T_c from chemical composition is especially attractive in this context [3]. Unlike properties that can be efficiently estimated using density functional theory, accurate first-principles prediction of T_c remains computationally challenging and often requires significant approximations. As a result, statistical learning models provide a complementary strategy for identifying promising candidate materials prior to experimental validation.

While prior studies have shown that ensemble methods outperform simple regression models for superconductivity prediction [3], the rapid development of modern boosting algorithms, neural networks, and ensemble stacking techniques motivates a systematic re-evaluation of existing benchmarks. It remains unclear how much performance improvement contemporary nonlinear models provide relative to linear baselines, whether advanced boosting methods surpass classical random forests, and how stable these flexible models are under cross-validation.

In this work, we conduct a comprehensive empirical comparison of linear regression, regularized linear models, random forests, gradient boosting, XGBoost, multi-layer perceptron neural networks, and stacked ensembles on the UCI Superconductivity dataset [3]. Our goal is to quantify the benefits of nonlinear modeling, evaluate the marginal gains from advanced ensemble strategies, and analyze the bias–variance tradeoffs that arise in flexible learners applied to materials data. By doing so, we establish a modern benchmark for compositional superconductivity prediction and provide practical insight into modeling strategies for structured scientific datasets.

II. Background

A. Materials Informatics and Superconductivity Prediction

Materials informatics has emerged as a powerful paradigm for accelerating materials discovery by integrating domain knowledge with data-driven learning. Instead of explicitly simulating quantum mechanical behavior for each candidate compound, machine learning models are trained to learn mappings between engineered descriptors and experimentally measured properties. This approach has been successfully applied to a wide range of materials properties, including formation energy, band gap, mechanical strength, and thermal conductivity.

In the context of superconductivity, predictive modeling focuses on estimating the superconducting critical temperature T_c. The mechanisms underlying superconductivity involve collective electronic interactions that depend on atomic composition, bonding environment, and electron–phonon coupling

effects. These relationships are highly nonlinear and involve complex interactions among elemental properties. Consequently, it is unlikely that T_c depends linearly on simple compositional statistics, motivating the use of flexible nonlinear function approximators.

B. UCI Superconductivity Dataset

The UCI Superconductivity dataset introduced by Hamidieh [3] contains 21,263 superconducting materials described by 81 engineered compositional features derived from elemental statistics. These descriptors include weighted averages, geometric means, entropies, ranges, and standard deviations of atomic properties such as atomic mass, first ionization energy, electron affinity, density, fusion heat, and valence electron counts.

Importantly, these features summarize composition without explicitly encoding crystal structure. As a result, the learning task is purely compositional, relying on statistical aggregation of elemental properties. The dataset provides a valuable benchmark for evaluating regression algorithms on structured tabular scientific data.

Figure 1 shows the distribution of Tc values, which is strongly right-skewed, with a large concentration of low-temperature materials and a long tail of higher-temperature superconductors.

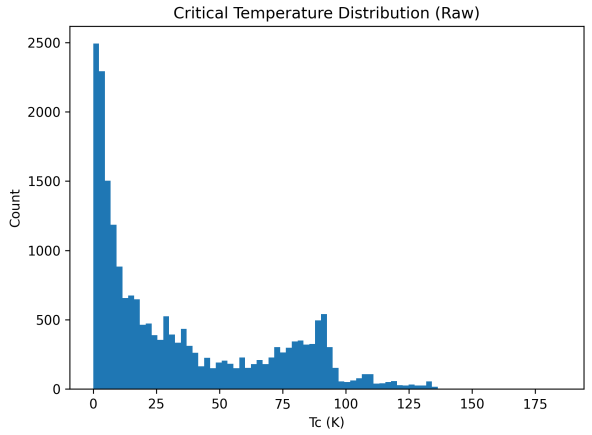

Fig. 1. Distribution of superconducting critical temperatures (T_c) in the dataset.

This skew introduces challenges for regression modeling, as minimizing squared error may bias models toward the dominant low-temperature regime.

C. Modeling Considerations

Linear regression models assume additive relationships between features and target values. While they offer interpretability and low variance, they cannot represent higher-order feature interactions unless explicitly engineered. In contrast, tree-based ensemble methods such as Random Forest and Gradient Boosting naturally capture nonlinear relationships and heterogeneous interactions across features. Boosting algorithms iteratively refine weak learners to reduce residual error, while bagging-based approaches such as Random Forest reduce variance through aggregation of decorrelated trees.

Neural networks provide another class of flexible nonlinear models capable of learning complex feature transformations. However, their performance on moderate-sized tabular datasets can vary depending on architecture design and regularization.

Given these modeling alternatives, systematic benchmarking is essential to determine which methods most effectively capture the compositional determinants of superconducting critical temperature and how they balance predictive accuracy with stability.

III. Methodology

A. Problem Formulation

Let $\mathbf{x}_i \in \mathbb{R}^d$ denote the engineered compositional feature vector for material i, and let $y_i = T_{c,i} \in \mathbb{R}$ denote its experimentally measured superconducting critical temperature. Given a dataset of n samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the objective is to learn a function

$$f_\theta : \mathbb{R}^d \to \mathbb{R} \tag{1}$$

parameterized by θ, such that predictions $\hat{y}_i = f_\theta(\mathbf{x}_i)$ minimize expected generalization error on unseen materials.

Model training is formulated as an empirical risk minimization problem:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, f_\theta(\mathbf{x}_i)\big), \tag{2}$$

where $\ell(\cdot)$ denotes a regression loss function. For all models considered in this study, optimization is based on squared error unless otherwise specified.

Model performance is evaluated using three complementary metrics. The Root Mean Squared Error (RMSE) is defined as

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2}, \tag{3}$$

where m denotes the number of validation samples. RMSE penalizes large errors more strongly due to squaring.

The Mean Absolute Error (MAE) is defined as

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i|, \tag{4}$$

which provides a more robust measure of average deviation.

Finally, we report the coefficient of determination,

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{m} (y_i - \bar{y})^2}, \tag{5}$$

where $\bar{y}$ is the mean of the validation targets. The R^2 metric quantifies the proportion of variance in T_c explained by the model.

B. Experimental Protocol

The dataset was randomly partitioned into 80% training and 20% validation subsets. All preprocessing operations were performed exclusively using training-set statistics to avoid data leakage. In particular, scale-sensitive models were standardized using the mean and variance computed on the training data, with identical transformations applied to validation samples. All experiments were implemented in Python using scikit-learn and XGBoost, and repeated over 10 random seeds.

In addition to single-split validation, five-fold cross-validation was performed to assess model stability and variance across different data partitions. In this setting, the dataset was divided into five disjoint folds; each fold was used once as validation while the remaining four folds were used for training. Reported cross-validation scores are summarized as mean $\pm$ standard deviation across folds.

C. Model Families

We evaluate a range of models spanning increasing representational complexity. The baseline model predicts the empirical mean of the training targets for all validation samples. This establishes a lower bound corresponding to zero explained variance.

Linear regression models estimate parameters by minimizing squared error under the assumption of an additive relationship between features and target. Ordinary Least Squares provides an unbiased estimator under classical assumptions, while Ridge and Lasso regression introduce ℓ_2 and ℓ_1 regularization respectively to reduce overfitting and encourage coefficient shrinkage [4], [5].

Tree-based ensemble methods are employed to capture nonlinear feature interactions. Random Forest constructs an ensemble of decision trees trained on bootstrapped subsets of the data with randomized feature selection at each split [6]. Predictions are obtained by averaging across trees, reducing variance relative to individual trees. Hyperparameter tuning was performed to optimize tree depth, number of estimators, and feature sampling strategies.

Gradient Boosting sequentially constructs trees that correct residual errors of previous models through additive stage-wise optimization [7]. A tuned variant was evaluated to assess the impact of learning rate, tree depth, and number of boosting stages. XGBoost, an optimized implementation of gradient boosting with regularization and second-order optimization, was included to evaluate the effect of more advanced boosting techniques [8].

A multi-layer perceptron (MLP) neural network was implemented with two hidden layers consisting of 128 and 64 units respectively, using ReLU activation functions [9], [10]. The network parameters were optimized using stochastic gradient-based methods to minimize mean squared error. This architecture enables learning of nonlinear transformations of compositional features without manual interaction engineering.

Finally, a stacked ensemble was constructed by training a meta-learner on predictions generated by top-performing base models. This approach seeks to combine complementary inductive biases and reduce generalization error through model aggregation [11].

Together, these models span a spectrum from low-variance linear estimators to highly flexible nonlinear learners, enabling systematic analysis of representational capacity, predictive accuracy, and stability in superconductivity modeling.

IV. Results

A. Overall Model Performance

Table I summarizes validation performance across all evaluated models. For models with multiple configurations (e.g., tuned variants), we report only the best-performing version.

TABLE I
VALIDATION PERFORMANCE (BEST CONFIGURATION PER MODEL).

Model	RMSE (K)	MAE (K)	R^2
Mean Baseline	33.93	29.10	-0.000
Linear Regression	17.38	13.21	0.738
Ridge Regression	17.40	13.22	0.737
Lasso Regression	17.48	13.29	0.735
Gradient Boosting	12.38	8.57	0.867
XGBoost	9.97	5.96	0.914
MLP Neural Network	10.49	6.63	0.904
Stacked Ensemble	9.31	5.22	0.925
Random Forest	**9.00**	**5.09**	**0.930**

The mean predictor yields no explanatory power, confirming that the dataset contains significant learnable structure. Linear regression explains approximately 74% of the variance in superconducting critical temperature, indicating that compositional descriptors encode substantial linear relationships.

However, nonlinear methods produce dramatic improvements. Random Forest achieves the strongest validation performance with $R^2 = 0.93$, reducing RMSE from 17.38 K (linear regression) to 9.00 K. This represents nearly a 48% reduction in RMSE relative to linear regression, highlighting the importance of nonlinear feature interactions. Figure 2 presents the parity plot for the Random Forest model, and Figure 3 graphs the prediction error distribution.

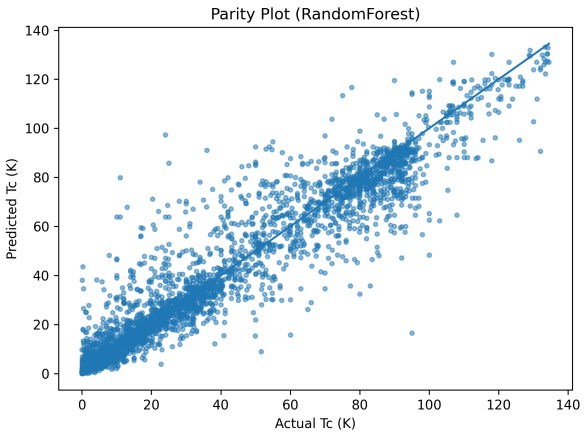

Fig. 2. Parity plot comparing predicted and actual superconducting critical temperatures for the Random Forest model. Points close to the diagonal indicate accurate predictions.

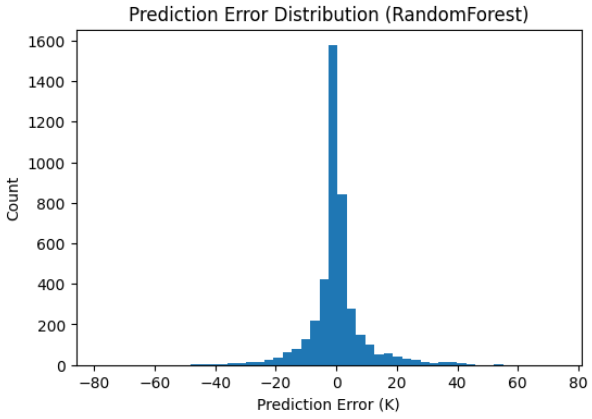

Fig. 3. Distribution of prediction errors for the Random Forest model. Most values are concentrated near zero, indicating accurate predictions for the majority of materials.

Boosting-based methods also perform strongly. XGBoost achieves $R^2 = 0.914$, outperforming classical Gradient Boosting, which attains $R^2 = 0.867$. The MLP neural network reaches $R^2 = 0.904$, demonstrating that neural architectures can effectively model nonlinear compositional relationships, though they do not surpass tree-based ensembles in this dataset.

The stacked ensemble achieves $R^2 = 0.925$, approaching Random Forest performance. This suggests that combining complementary learners yields modest additional gains but does not fundamentally outperform the strongest individual tree ensemble.

Overall, tree-based ensemble methods consistently achieve the highest predictive accuracy, indicating that heterogeneous, nonlinear interactions among elemental statistics govern superconducting critical temperature.

B. Cross-Validation Stability

To assess model robustness, five-fold cross-validation was performed. Table II reports mean R^2 and standard deviation across folds.

TABLE II
FIVE-FOLD CROSS-VALIDATION R^2 SCORES (MEAN $\pm$ STD).

Model	R^2 (mean $\pm$ std)
Linear Regression	0.69 ± 0.12
Random Forest	0.52 ± 0.17
Gradient Boosting	0.55 ± 0.18
XGBoost	0.53 ± 0.18
MLP Neural Network	0.60 ± 0.23
Stacked Ensemble	0.53 ± 0.18

While nonlinear models achieve higher peak performance on the held-out validation split, cross-validation reveals increased fold-to-fold variability relative to linear regression. The larger standard deviations observed for boosting and neural models indicate greater sensitivity to data partitioning.

This behavior reflects the classical bias–variance tradeoff. Linear regression exhibits higher bias but lower variance, producing more stable cross-validation scores. In contrast, flexible nonlinear models reduce bias by capturing complex feature interactions, but incur higher variance across folds.

The disparity between single-split validation performance and cross-validation averages suggests that dataset partitioning meaningfully influences performance estimates. Given the right-skewed distribution of T_c and heterogeneous feature effects, flexible learners may adapt strongly to specific training subsets.

C. Interpretation

Across all experiments, nonlinear ensemble methods provide substantial gains over linear baselines, confirming that superconducting critical temperature depends on complex, non-additive interactions among compositional descriptors. Random Forest emerges as the most reliable high-performing model, while boosting and stacking offer competitive alternatives.

However, cross-validation analysis emphasizes that increased representational capacity comes with higher variability. These findings highlight the importance of robust validation protocols when applying flexible machine learning models to scientific datasets with skewed distributions and heterogeneous structure.

V. DISCUSSION

The results of this study demonstrate that superconducting critical temperature is governed by complex, nonlinear relationships among compositional descriptors. While linear regression explains approximately 74% of the variance in T_c, the substantial reduction in prediction error achieved by ensemble tree methods indicates that additive feature assumptions are insufficient to capture the underlying structure of the data. The nearly 48% reduction in RMSE achieved by Random Forest relative to linear regression underscores the importance of modeling higher-order interactions among elemental statistics.

Tree-based ensembles consistently outperform both linear models and neural networks in this dataset. Random Forest, in particular, achieves the strongest validation performance. Its success can be attributed to several factors. First, decision trees naturally partition feature space into regions that capture heterogeneous interactions without requiring explicit feature engineering. Second, bagging reduces variance by averaging across decorrelated trees, providing stability in moderate-sized tabular datasets. These properties make Random Forest especially well-suited to compositional materials data, where nonlinear interactions and feature heterogeneity are expected but spatial structure is absent.

Boosting-based methods such as XGBoost also achieve strong performance, confirming that additive stage-wise tree optimization effectively captures residual nonlinear structure. However, boosting methods exhibit greater sensitivity to hyperparameter configuration and data partitioning. The stacked ensemble provides modest additional improvement, suggesting that complementary inductive biases among models can yield incremental gains, though the improvement is not dramatic relative to the strongest individual ensemble.

Neural networks achieve competitive but slightly lower performance than tree ensembles. This outcome is consistent with broader findings in structured tabular learning, where tree-based methods often outperform deep networks unless very large datasets are available. The moderate size of the dataset (approximately 21,000 samples) and the absence of spatial or sequential structure likely limit the relative advantage of neural architectures in this setting.

Cross-validation analysis reveals an important tradeoff. Although nonlinear models achieve higher peak validation performance, they exhibit larger fold-to-fold variability compared to linear regression. This behavior reflects the classical bias–variance tradeoff: linear models incur higher bias but lower variance, while flexible learners reduce bias at the cost of increased variance. The skewed distribution of T_c and heterogeneous feature interactions may further amplify this variability, as different folds may contain differing proportions of high-temperature materials.

From a materials informatics perspective [2], these findings reinforce the hypothesis that superconducting critical temperature emerges from complex combinations of elemental properties rather than simple linear trends. The effectiveness of tree-based ensembles suggests that interaction effects—potentially reflecting coupled electronic and bonding phenomena—play a dominant role in determining T_c when represented through compositional statistics.

VI. Conclusion

In this work, we conducted a comprehensive empirical comparison of linear regression, regularized linear models, tree-based ensembles, boosting methods, neural networks, and stacked ensembles for predicting superconducting critical temperature from compositional descriptors.

Nonlinear ensemble methods substantially outperform linear baselines, with Random Forest achieving the strongest validation performance ($R^2 = 0.93$). Boosting methods and stacked ensembling provide competitive alternatives, while neural networks demonstrate effective but slightly lower performance in this tabular setting. Cross-validation analysis highlights increased variance for flexible models, emphasizing the importance of careful validation and model selection.

Collectively, these results confirm that nonlinear feature interactions are essential for accurate modeling of superconducting critical temperature. The study establishes an updated benchmark for compositional superconductivity prediction using modern ensemble and boosting techniques.

Future work may investigate log-transformed target modeling to address distributional skew, incorporate uncertainty quantification to better assess predictive confidence, and explore physics-informed feature representations or hybrid models that integrate domain knowledge with data-driven learning. Such approaches may further improve robustness and generalization in materials property prediction.

References

[1] J. G. Bednorz and K. A. Müller, "Possible high tc superconductivity in the ba-la-cu-o system," *Zeitschrift für Physik B*, vol. 64, pp. 189–193, 1986.

[2] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," *Nature*, vol. 559, pp. 547–555, 2018.

[3] K. Hamidieh, "A data-driven statistical model for predicting the critical temperature of a superconductor," *Computational Materials Science*, vol. 154, pp. 346–354, 2018.

[4] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B*, vol. 58, pp. 267–288, 1996.

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[7] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 2001.

[8] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[11] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992.