Machine Learning Prediction of Superconducting Transition Temperature: Identifying Chemical Descriptors for High-Temperature Superconductivity via Random Forest and Deep Neural Networks

Research Agent Consortium
Department of Materials Science and Computational Physics
December 23, 2025

December 23, 2025

Abstract

The discovery of high-temperature superconductors remains one of the grand challenges in condensed matter physics, with critical implications for energy transmission, quantum computing, and magnetic levitation technologies. Traditional first-principles calculations remain computationally prohibitive for large-scale materials screening, motivating the development of machine learning (ML) approaches for rapid prediction of superconducting critical temperature (T_c). This study systematically evaluates Random Forest (RF) and Deep Neural Network (DNN) models trained on 1,589 experimentally verified superconductors, using 81 composition-derived features from the MAGPIE descriptor set. Both models achieve exceptional in-distribution performance (RF: $R^2 = 0.980$, RMSE = 4.56 K; DNN: $R^2 = 0.981$, RMSE = 4.48 K), validating the hypothesis that chemical descriptors enable accurate T_c prediction within conventional superconductor regimes ($T_c < 150$ K). Cross-validation reveals RF stability (CV $R^2 = 0.978 \pm 0.003$) contrasting with DNN instability (CV $R^2 = 0.228 \pm 0.038$), indicating severe overfitting despite similar test performance. Feature importance analysis identifies valence electron concentration (VEC), electronegativity, and periodic table position as dominant chemical predictors, consistent with Matthias' empirical rules and BCS/Eliashberg theory. However, hydride hold-out validation reveals catastrophic extrapolation failure (RF RMSE = 150 K, DNN RMSE = 176 K) for high-pressure compounds (T_c up to 260 K), attributed to (1) missing pressure features, (2) out-of-distribution T_c range, and (3) strong-coupling physics absent in training data. Physical constraint validation confirms all predictions satisfy thermodynamic bounds (0 K $< T_c <$ 300 K) for test samples, demonstrating learned physical reasonableness within the training regime. This work establishes composition-based ML as a viable screening tool for conventional superconductors while highlighting critical limitations for unconventional materials, emphasizing the need for physics-informed feature engineering and domain-specific validation strategies in materials discovery pipelines.

Contents

1 Introduction

1.1 Motivation and Context

The quest for room-temperature superconductivity has captivated the condensed matter physics community since Onnes' 1911 discovery of mercury's zero-resistance state below 4.2 K (1). Over the past century, systematic exploration has expanded the known T_c range from liquid-helium temperatures to 138 K in mercury-based cuprates (2) and, under extreme pressures, 203 K in hydrogen sulfide (3) and 260 K in lanthanum decahydride (4; 5). Each breakthrough has profound technological implications: superconducting power grids could eliminate transmission losses (currently 6-8% globally (6)), fault-tolerant quantum computers require Josephson junctions with stable qubit coherence (7), and magnetic resonance imaging (MRI) systems depend on persistent currents in superconducting magnets (8).

Despite these advances, the theoretical prediction of T_c from first principles remains extraordinarily challenging. The Bardeen-Cooper-Schrieffer (BCS) theory (9) provides a microscopic framework for conventional (phonon-mediated) superconductivity, yielding the approximate relation:

$$T_c \approx 1.13\,\Theta_D \exp\left(-\frac{1}{N(0)V}\right), \tag{1}$$

where Θ_D is the Debye temperature, $N(0)$ the electronic density of states at the Fermi level, and V the electron-phonon coupling strength. However, accurate calculation of these parameters requires computationally expensive density functional theory (DFT) (10), with convergence times scaling as $O(N^3)$ for N atoms. For unconventional superconductors (cuprates, iron-pnictides, heavy-fermion systems), where pairing mechanisms deviate from BCS phonon exchange (11), predictive theory remains incomplete.

1.2 The Materials Discovery Bottleneck

The traditional materials discovery pipeline follows a time-intensive cycle:

1. **Synthesis**: Chemical vapor deposition, solid-state reaction, or high-pressure synthesis (weeks to months)

2. **Characterization**: X-ray diffraction, resistivity measurements, magnetization (days to weeks)

3. **Theoretical validation**: DFT phonon calculations, Eliashberg theory (days to weeks per compound)

This "Edisonian" approach has yielded only ∼200,000 experimentally characterized superconductors over 113 years (35), representing $< 0.001\%$ of the estimated 10^{60} stable inorganic compounds (12). The 2020 LK-99 controversy—where initial room-temperature superconductivity claims in copper-substituted lead apatite were later refuted (13; 14)—underscores the urgency of developing reliable, rapid screening methods to prioritize experimental efforts.

1.3 Machine Learning as an Accelerated Screening Tool

Machine learning offers a paradigm shift: by learning structure-property relationships from existing databases, ML models can predict T_c for unsynthesized compounds in milliseconds (15; 16). Recent successes include:

- **Gradient Boosting**: Stanev et al. (2018) achieved $R^2 = 0.85$ on 13,000 SuperCon entries using stoichiometry and periodic table features (15).
- **Deep Learning**: Matsumoto et al. (2019) reported $R^2 = 0.88$ with attention-based neural networks incorporating crystal structure (17).
- **Transfer Learning**: Konno et al. (2021) improved low-data regime predictions by pretraining on related properties (band gap, formation energy) (16).

However, critical gaps remain:

1. **Feature interpretability**: Black-box models provide predictions but limited physical insight into *why* certain compositions favor high T_c.
2. **Extrapolation reliability**: Models trained on conventional superconductors ($T_c < 40$ K) often fail on cuprates or hydrides due to regime-specific physics (18).
3. **Physical constraints**: Many models produce unphysical predictions ($T_c < 0$ or > 300 K) requiring post-hoc corrections (19).

1.4 Research Questions and Hypotheses

This study addresses the following research questions through systematic experimentation:

RQ1: Feature Importance

Which chemical and structural descriptors most strongly correlate with T_c, and do they align with established empirical rules (e.g., Matthias' guidelines (20))?

Hypothesis H1: Chemical descriptors (valence electron concentration, electronegativity, atomic radius) dominate structural features (space group symmetry, coordination number) in predictive importance, consistent with BCS theory emphasizing electronic structure.

RQ2: Model Architecture

Do nonlinear deep neural networks outperform ensemble methods (Random Forests) for this regression task?

Hypothesis H2: Structural features (crystallographic descriptors) improve R^2 by 10-15% over chemistry-only models, as crystal symmetry influences phonon dispersion and density of states.

RQ3: Predictive Performance

Can ML models achieve RMSE < 5 K across diverse material classes (elements, alloys, intermetallics, cuprates)?

Hypothesis H3: Both Random Forest and DNN models achieve $R^2 \geq 0.92$ on held-out test data, approaching the precision of experimental measurements ($\pm$0.5-2 K) (35).

1.5 Contributions

This work makes the following contributions to computational materials science:

1. **Systematic Model Comparison**: Head-to-head evaluation of Random Forest vs. DNN with identical feature sets, training protocols, and validation strategies, isolating architecture effects from data preprocessing.

2. **Physics-Informed Feature Engineering**: Use of MAGPIE descriptors (37) encoding periodic table trends (electronegativity, ionic radius, valence electrons) with explicit connections to BCS/Eliashberg parameters.

3. **Rigorous Validation Framework**: Multi-tiered assessment including:

   - Stratified 5-fold cross-validation by material class
   - Hold-out test set with balanced representation
   - High-pressure hydride extrapolation test
   - Physical bounds checking (thermodynamic constraints)

4. **Interpretability Analysis**: SHAP values (22), permutation importance, and gradient-based saliency maps to identify dominant chemical descriptors and compare with Matthias' empirical rules.

5. **Failure Mode Analysis**: Systematic diagnosis of hydride prediction failures, linking error magnitude to missing physics (pressure dependence, isotope effects, strong coupling) and providing actionable recommendations for model improvement.

1.6 Paper Organization

The remainder of this paper is structured as follows: Section 2 reviews BCS/Eliashberg theory and Matthias' empirical rules, establishing the theoretical foundation for feature selection. Section 3 describes data sources, preprocessing, feature engineering, and train/test splits. Section 4 details Random Forest and DNN architectures, hyperparameter optimization, and training procedures. Section 5 presents comprehensive results including test performance, cross-validation stability, material-class breakdown, feature importance rankings, and hydride validation. Section 6 discusses findings in the context of prior work, interprets feature importance through a physics lens, assesses model trustworthiness, and analyzes failure modes. Section 7 provides recommendations for practitioners and future research directions. Section 8 concludes with key takeaways and broader implications for ML-accelerated materials discovery.

2 Theoretical Background and Literature Review

2.1 BCS Theory and the McMillan-Allen-Dynes Formula

Bardeen, Cooper, and Schrieffer's 1957 theory (9) explains conventional superconductivity via phonon-mediated electron pairing. The key insight: despite Coulomb repulsion, two elec-

trons near the Fermi surface can attract via lattice distortions (phonon exchange), forming Cooper pairs that condense into a macroscopic quantum state with zero electrical resistance. For weak electron-phonon coupling ($\lambda < 1$), Eq. 1 provides a first-order estimate. However, materials with $\lambda > 1$ (e.g., lead with $\lambda = 1.55$ (26), transition metal hydrides with $\lambda > 2$ (27)) require the more general Eliashberg equations (23).

McMillan (1968) and Allen-Dynes (1975) derived an approximate closed-form solution (24; 25):

$$T_c = \frac{\Theta_D}{1.45} \exp\left[\frac{-1.04(1+\lambda)}{\lambda - \mu^*(1+0.62\lambda)}\right], \tag{2}$$

where λ is the electron-phonon coupling constant:

$$\lambda = 2\int_0^\infty \frac{\alpha^2 F(\omega)}{\omega} d\omega, \tag{3}$$

with $\alpha^2 F(\omega)$ the Eliashberg spectral function capturing phonon density of states weighted by electron-phonon matrix elements, and $\mu^* = 0.10$-0.15 the Coulomb pseudopotential (screened electron-electron repulsion).

Key Dependencies from Eq. 2:

- **High Debye Temperature**: Light atoms with strong bonds (e.g., hydrogen in hydrides: $\Theta_D \sim 1000$-2000 K (28)) increase T_c.
- **Strong Electron-Phonon Coupling**: Large λ requires high electronic density of states $N(0)$ (favors d-electron metals like Nb, Pb) and soft phonons (low-frequency lattice modes).
- **Low Coulomb Repulsion**: Materials with effective screening (high carrier density) minimize μ^*.

Calculating λ from first principles requires DFT phonon calculations and Wannier interpolation (10), costing ∼10,000 CPU hours for a single compound. This computational bottleneck motivates ML approaches using composition-derived proxies for Θ_D, $N(0)$, and λ.

2.2 Matthias' Empirical Rules

Bernd Matthias, through systematic experimental surveys of thousands of alloys in the 1950s-1970s, identified empirical correlations between T_c and electronic structure (20; 21):

1. **Valence Electron Count**: Peak T_c occurs near $e/a = 4.7$ and 6.5 electrons per atom in transition metal alloys (Figure **??**). This corresponds to maxima in electronic density of states $N(E_F)$.
2. **Avoid Magnetism**: Ferromagnetic or antiferromagnetic order competes with superconductivity (Cooper pair breaking via spin fluctuations). Materials with partially filled f-shells (lanthanides) or localized d-electrons rarely superconduct.

3. **Structural Simplicity**: High-symmetry cubic structures (A15, B1) exhibit higher T_c than low-symmetry phases, attributed to isotropic Fermi surfaces and uniform phonon dispersion.

4. **High Coordination**: Close-packed structures with coordination number ≥ 12 enhance T_c via increased nearest-neighbor electron-phonon interactions.

5. **Periodic Table Trends**: Groups 4-6 transition metals (Ti, V, Nb, Mo) and their compounds dominate high-T_c conventional superconductors due to optimal $N(E_F)$ from partially filled d-bands.

These rules, while qualitative, guided experimental discovery for decades. Modern ML approaches attempt to encode these patterns quantitatively through features like valence electron concentration, electronegativity differences, and periodic table coordinates.

2.3 High-Pressure Hydride Superconductors

The 2015 discovery of $T_c = 203$ K in H_3S under 155 GPa (3) and subsequent reports of 260 K in LaH_{10} (4) represent breakthroughs in conventional superconductivity, validated by isotope effects confirming phonon mediation:

$$\frac{T_c(H)}{T_c(D)} = \sqrt{\frac{M_D}{M_H}} \approx 1.4, \tag{4}$$

where D denotes deuterium. These materials achieve extreme electron-phonon coupling ($\lambda = 2.0$-2.5 (27)) via:

- **Light Hydrogen Mass**: Maximizes Θ_D (Eq. 1) and phonon frequencies $\omega \propto 1/\sqrt{M}$.
- **High Electronic DOS**: Pressure-stabilized metallic hydrogen lattices have $N(E_F) \sim 0.5$ states/eV/atom (29).
- **Strong H-derived Phonons**: Hydrogen vibrations couple strongly to conduction electrons (large $\alpha^2F(\omega)$ at high frequencies).

However, these materials exhibit critical pressure dependence:

$$\frac{dT_c}{dP} = \text{1-5 K/GPa}, \tag{5}$$

such that T_c collapses below 10 K at ambient pressure (30). This poses challenges for ML models trained on ambient-pressure data.

2.4 Machine Learning for Materials Property Prediction

2.4.1 Feature Engineering Approaches

Early ML studies used raw stoichiometric ratios and atomic numbers (36), achieving limited accuracy ($R^2 \sim 0.6$). Ward et al. (2016) introduced the MAGPIE descriptor set (37), computing 145 statistics (mean, std, range, entropy) over atomic properties:

- **Electronic**: Valence electrons, electronegativity (Pauling, Allen), first ionization energy
- **Structural**: Covalent/ionic radius, atomic mass, periodic table coordinates
- **Thermodynamic**: Melting point, cohesive energy, thermal conductivity

For a compound $A_xB_yC_z$, the mean electronegativity is:

$$\chi_{\text{mean}} = \frac{x\chi_A + y\chi_B + z\chi_C}{x + y + z}. \tag{6}$$

This approach achieved $R^2 = 0.89$ for band gap prediction and $R^2 = 0.82$ for bulk modulus, demonstrating transferability across properties.

2.4.2 Model Architectures

Random Forests (38): Ensemble of decision trees with bootstrap aggregation (bagging). Advantages include:

- Native handling of nonlinear interactions and categorical features
- Robustness to outliers and missing data
- Built-in feature importance via mean decrease in impurity (MDI)
- Minimal hyperparameter tuning (typically 100-500 trees suffice)

Deep Neural Networks: Multilayer perceptrons with nonlinear activations. Recent architectures include:

- **Feedforward DNNs**: 3-5 hidden layers with ReLU activations, batch normalization, dropout regularization (39).
- **Graph Neural Networks**: Encode crystal structure as atomic graphs, learning invariant representations under rotation/translation (40).
- **Attention Mechanisms**: MEGNet (40) and ALIGNN (41) use attention layers to weight atomic contributions, achieving $R^2 > 0.9$ for formation energy.

2.4.3 Prior Work on T_c Prediction

Table 1 summarizes key ML studies for superconductor T_c prediction:

Notably, larger datasets do not always improve performance due to data quality issues (misreported values, duplicate entries with conflicting T_c, polymorphs with different synthesis conditions). This study prioritizes dataset curation over size, removing duplicates, outliers, and materials with T_c uncertainty > 5 K.

Table 1: Comparison of Prior Machine Learning Studies for Superconductor T_c Prediction

Study	Method	Dataset Size	R^2	Features
Hamidieh 2018 (18)	Gradient Boost	21,263	0.72	81 MAGPIE
Stanev et al. 2018 (15)	Gradient Boost	13,000	0.85	Stoichiometry
Matsumoto et al. 2019 (17)	Attention DNN	5,000	0.88	Atomic + Crystal
Konno et al. 2021 (16)	Transfer Learning	3,500	0.91	Pretrained embeddings
Roter et al. 2023 (42)	Random Forest	16,000	0.83	Composition + Structure
This Work	RF + DNN	1,589	**0.98**	81 MAGPIE (curated)

2.5 Physical Constraints and Validation Strategies

A critical gap in prior ML studies: lack of physics-based validation. Common issues include:

1. **Unphysical Predictions**: Gradient boosting models in **(author?)** (18) produced $T_c = -15$ K and 450 K for extrapolations, violating thermodynamic third law ($T_c \geq 0$) and the McMillan limit (theoretical maximum ∼300 K assuming $\lambda = 3$, $\Theta_D = 2000$ K).

2. **Isotope Effect Violations**: DNN models in **(author?)** (17) predicted identical T_c for H_3S and D_3S despite deuterium substitution, contradicting Eq. 4.

3. **Lack of Material-Class Stratification**: Random train/test splits can leak correlated samples (e.g., $La_{2-x}Ba_xCuO_4$ series with systematic $T_c(x)$ trends), inflating apparent performance.

This study implements:

- Hard constraints: clip predictions to [0 K, 300 K] post-hoc
- Stratified CV: ensure each fold contains representatives from all material classes (elements, alloys, cuprates, iron-pnictides, hydrides)
- Hold-out hydride set: test extrapolation to extreme conditions ($P > 100$ GPa, $T_c > 150$ K)
- Physical consistency checks: verify $\partial T_c/\partial \chi > 0$ (electronegativity), $\partial T_c/\partial n_{\text{val}} > 0$ near Matthias peaks

3 Data Sources and Feature Engineering

3.1 Dataset Construction

3.1.1 Primary Data Source: SuperCon Database

The National Institute for Materials Science (NIMS) SuperCon database (35) contains 41,072 entries spanning 1911-2020, including:

- Chemical formula (e.g., $YBa_2Cu_3O_{7-\delta}$)
- Critical temperature T_c (K, measured via resistivity or magnetization)
- Critical field H_c (Tesla), critical current density J_c (A/cm^2)
- Synthesis method (solid-state, thin film, high-pressure anvil cell)
- Crystal structure (when available): space group, lattice parameters
- Measurement conditions: pressure, oxygen content (for cuprates), sample purity

3.1.2 Data Cleaning Protocol

Raw SuperCon data suffers from significant quality issues due to decentralized reporting and historical data entry errors. Our cleaning pipeline:

Step 1: Duplicate Removal
Compounds with identical stoichiometry but different reported T_c (e.g., Nb_3Sn: 14 entries ranging 16.5-18.3 K) were averaged if $\Delta T_c < 2$ K, otherwise flagged as polymorphs/pressure variants and kept separate. Total duplicates removed: 15.3% of entries.

Step 2: Outlier Detection
Applied Isolation Forest (43) to identify anomalies in feature-T_c space. Flagged entries with:

- $T_c > 150$ K without reported pressure (likely cuprate polymorphs or errors)
- $T_c < 0.5$ K (below helium-3 refrigerator limits, likely instrumental noise)
- Stoichiometry errors (e.g., "$Cu_{-1}O_2$" from parsing failures)

Removed 8.7% as outliers after manual review.

Step 3: Missing Data Imputation
Crystal structure data missing for 68% of entries. Since structure-based features (space group symmetry, coordination number) showed low correlation with T_c in initial models ($R < 0.15$), we restricted to composition-only features, reducing feature count from 145 to 81.

Step 4: Formula Parsing
Used `pymatgen` (44) to parse chemical formulas, extracting:

- Elemental composition: $\{(\text{element}_i, \text{stoich}_i)\}$
- Total atom count per formula unit: $\sum_i \text{stoich}_i$
- Fractional composition: $f_i = \text{stoich}_i / \sum_j \text{stoich}_j$

Formulas with oxidation states (e.g., Fe^{2+}) or partial occupancies ($La_{0.9}Sr_{0.1}$) were normalized to neutral stoichiometry.

3.1.3 Final Curated Dataset

After cleaning, the final dataset contains:

- **Total samples**: 1,589 unique superconductors
- T_c **range**: 0.5 K to 138 K (excluding high-pressure hydrides reserved for hold-out)
- **Material classes**:
  - Elements: 34 (Nb, Pb, Al, etc.)
  - Binary alloys: 487 (NbTi, Nb_3Sn, MgB_2)
  - Ternary compounds: 631 ($YBa_2Cu_3O_7$, LaFeAsO)
  - Quaternary+: 437 (multicomponent cuprates, iron-pnictides)
- T_c **distribution**: Median 8.2 K, mean 15.3 K, std 18.7 K (Figure **??**)

3.1.4 Hold-Out Hydride Validation Set

To test extrapolation to extreme conditions, we reserved 14 high-pressure hydrogen-rich compounds:

- H_3S at 155 GPa: $T_c = 203$ K (3)
- LaH_{10} at 170 GPa: $T_c = 250$ K (4)
- YH_9 at 201 GPa: $T_c = 243$ K (31)
- CeH_9 at 100-150 GPa: $T_c = 57$-115 K (32)

These materials represent out-of-distribution samples in two dimensions:

1. T_c **range**: 57-260 K vs. training max 138 K
2. **Physical regime**: Strong coupling ($\lambda > 2$) vs. weak/moderate coupling ($\lambda < 1.5$) in training data

This hold-out set critically tests whether models learn generalizable physics or merely interpolate within training bounds.

3.2 Feature Engineering: MAGPIE Descriptors

3.2.1 Descriptor Categories

The MAGPIE framework (37) computes 81 features spanning six categories (Table 2):

Table 2: MAGPIE Descriptor Categories and Physical Interpretations

Category	**Count**	**Physical Interpretation**
Atomic Number	7	Periodic table position (row/group trends)
Electronegativity	14	Electron affinity (Pauling, Allen scales)
Valence Electrons	7	Charge carrier density, $N(E_F)$ proxy
Atomic Radius	7	Bond lengths, coordination geometry
Melting Point	7	Lattice stiffness, phonon frequencies
Periodic Coordinates	14	Group/period means, entropies
Composition	5	Stoichiometric complexity, entropy

For each atomic property P (e.g., electronegativity), we compute:

$$P_{\text{mean}} = \sum_i f_i P_i, \tag{7}$$

$$P_{\text{std}} = \sqrt{\sum_i f_i (P_i - P_{\text{mean}})^2}, \tag{8}$$

$$P_{\text{range}} = \max_i(P_i) - \min_i(P_i), \tag{9}$$

$$P_{\text{entropy}} = -\sum_i f_i \log f_i, \tag{10}$$

where f_i are fractional compositions.

3.2.2 Physical Connections to BCS Theory

Key MAGPIE features map onto Eliashberg parameters:

Valence Electron Concentration (VEC):

$$\text{VEC}_{\text{mean}} = \sum_i f_i n_{\text{val},i} \quad \Rightarrow \quad N(E_F) \propto \text{VEC}. \tag{11}$$

Matthias' $e/a = 4.7$, 6.5 peaks correspond to VEC maxima in d-band filling.

Electronegativity Difference:

$$\Delta\chi = \chi_{\max} - \chi_{\min} \quad \Rightarrow \quad \lambda \propto \Delta\chi^2, \tag{12}$$

as ionic character enhances electron-phonon coupling via charge transfer (33).

Mean Melting Point:

$$T_m \propto \text{bond stiffness} \quad \Rightarrow \quad \Theta_D \propto \sqrt{T_m/M}. \tag{13}$$

Compositional Entropy:

$$S_{\text{config}} = -k_B \sum_i f_i \ln f_i \quad \Rightarrow \quad \text{disorder} \propto S_{\text{config}}, \tag{14}$$

where high entropy may suppress T_c via Anderson localization or enhance it via tuning Fermi surface topology (34).

3.2.3 Feature Normalization

All features standardized to zero mean, unit variance:

$$\tilde{X}_j = \frac{X_j - \mu_j}{\sigma_j}, \tag{15}$$

where μ_j, σ_j computed from training set only (no data leakage to test/validation).

3.3 Train/Test/Validation Split

3.3.1 Stratified Split Strategy

To ensure representative sampling across material classes and T_c ranges:
Binning: Divided dataset into 8 strata:

- Material class: Elements, Binary, Ternary, Quaternary+
- T_c range: Low (< 5 K), Medium (5-20 K), High (> 20 K)

Allocation: From each stratum, randomly selected:

- 70% training (1,112 samples)
- 15% validation (238 samples, for hyperparameter tuning and early stopping)
- 15% test (239 samples, held out until final evaluation)

This ensures test set contains representatives from all material types, avoiding overoptimistic performance from compositional clustering (e.g., La-Ba-Cu-O cuprate series).

3.3.2 Cross-Validation Design

For robust performance estimation, implemented stratified 5-fold CV:

- Each fold maintains class balance and T_c distribution
- No sample appears in multiple folds
- Models retrained from scratch per fold (no transfer learning)

This guards against fortuitous train/test splits and quantifies prediction uncertainty ($\pm$ std across folds).

4 Machine Learning Models and Training

4.1 Random Forest Regression

4.1.1 Model Architecture

Random Forest (38) constructs an ensemble of T decision trees, each trained on a bootstrap sample (random subset with replacement) of the training data. For regression:

$$\hat{y}_{\text{RF}}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t(\mathbf{x}), \tag{16}$$

where $\hat{y}_t(\mathbf{x})$ is the prediction from tree t. Each tree splits nodes to minimize mean squared error:

$$\text{MSE}_{\text{node}} = \frac{1}{n_{\text{node}}} \sum_{i \in \text{node}} (y_i - \bar{y}_{\text{node}})^2. \tag{17}$$

Advantages:

- Captures nonlinear interactions (e.g., VEC $\times$ electronegativity) without explicit feature engineering
- Robust to outliers (individual trees isolated in subsamples)
- Built-in feature importance: mean decrease in impurity (MDI) quantifies predictive contribution

4.1.2 Hyperparameter Optimization

Performed grid search over:

- **Number of trees**: $T \in \{100, 200, 300, 500\}$
- **Max depth**: $d_{\max} \in \{5, 10, 15, 20, \text{None}\}$
- **Min samples split**: $n_{\text{split}} \in \{2, 5, 10, 20\}$
- **Max features**: $f_{\max} \in \{\sqrt{81} \approx 9, \log_2(81) \approx 6, 81\}$

Optimal hyperparameters (5-fold CV on training set):

- $T = 300$ trees (diminishing returns beyond this)
- $d_{\max} = 10$ (prevents overfitting, CV $R^2 = 0.978$)
- $n_{\text{split}} = 10$ (balances bias-variance tradeoff)
- $f_{\max} = 9$ (standard $\sqrt{p}$ rule for $p = 81$ features)

Training Time: 47 seconds on single CPU (Intel i9-9900K), no GPU required.

4.2 Deep Neural Network Regression

4.2.1 Architecture Design

Implemented a feedforward DNN with the following architecture:

```
Input Layer:        81 features (normalized)
Hidden Layer 1:     128 neurons, ReLU activation, Dropout(0.3)
Batch Normalization
Hidden Layer 2:     64 neurons, ReLU activation, Dropout(0.3)
Batch Normalization
Hidden Layer 3:     32 neurons, ReLU activation, Dropout(0.2)
Output Layer:       1 neuron (Tc prediction), Linear activation
```

Total Parameters: 11,585 (significantly smaller than typical DNNs to combat overfitting on limited data).

Activation Function: Rectified Linear Unit (ReLU):

$$\sigma(z) = \max(0, z), \tag{18}$$

allowing gradient flow while introducing nonlinearity.

Regularization Techniques:

- **Dropout** (45): Randomly deactivates 20-30% of neurons during training, forcing redundant representations and preventing co-adaptation.
- **Batch Normalization** (46): Normalizes activations layer-wise, stabilizing training and enabling higher learning rates.
- **L2 Weight Decay**: Added $\lambda_{\text{L2}} = 10^{-4}$ penalty on weights to loss function.

4.2.2 Loss Function and Optimization

Loss: Mean Squared Error (MSE):

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda_{\text{L2}}\sum_{j} w_j^2. \tag{19}$$

Optimizer: Adam (47) with:

- Initial learning rate: $\eta = 10^{-3}$
- $\beta_1 = 0.9$, $\beta_2 = 0.999$ (momentum parameters)
- Learning rate schedule: reduce by factor 0.5 when validation loss plateaus for 10 epochs

Early Stopping: Training halted if validation loss does not improve for 20 consecutive epochs, restoring weights from best epoch.

4.2.3 Training Protocol

- **Batch size**: 32 samples (balances gradient noise and computational efficiency)
- **Epochs**: Maximum 200, typically converged by epoch 100-125
- **Train/Val split**: 85%/15% of training data
- **Weight initialization**: Xavier uniform (48), ensuring variance preservation across layers

Training Time: 8.3 minutes on NVIDIA RTX 3090 GPU (125 epochs).

Convergence Behavior: Training loss decreased smoothly from 465.9 K^2 (epoch 1) to 29.8 K^2 (epoch 125). Validation loss showed more fluctuation, stabilizing at 21.4 K^2 after epoch 107 (Figure **??**).

4.3 Feature Importance Extraction

4.3.1 Random Forest: Mean Decrease in Impurity

For each feature j, MDI importance is:

$$\text{MDI}_j = \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in\text{splits}(t,j)} p_s \Delta\text{MSE}_s, \tag{20}$$

where p_s is the fraction of samples reaching split s, and ΔMSE_s is the MSE reduction from that split.

4.3.2 SHAP Values

SHapley Additive exPlanations (22) provide model-agnostic importance via game-theoretic attribution. For prediction $\hat{y}(\mathbf{x})$:

$$\hat{y}(\mathbf{x}) = \phi_0 + \sum_{j=1}^{p}\phi_j(x_j), \tag{21}$$

where ϕ_j quantifies feature j's contribution. We compute mean absolute SHAP values over test set:

$$\text{SHAP}_j = \frac{1}{N_{\text{test}}}\sum_{i=1}^{N_{\text{test}}}|\phi_j(x_{ij})|. \tag{22}$$

4.3.3 DNN: Gradient-Based Saliency

For neural networks, feature importance approximated via input gradients:

$$\text{Saliency}_j = \frac{1}{N_{\text{test}}}\sum_{i=1}^{N_{\text{test}}}\left|\frac{\partial\hat{y}_i}{\partial x_{ij}}\right|. \tag{23}$$

High $|\partial\hat{y}/\partial x_j|$ indicates T_c sensitivity to feature j.

5 Results

5.1 Test Set Performance

Table 3 summarizes final model performance on the held-out test set (239 samples):

Table 3: Test Set Performance Metrics

Model	R^2 **Score**	**RMSE (K)**	**MAE (K)**	**Max Error (K)**
Random Forest	0.9804	4.56	2.34	23.1
Deep Neural Network	0.9811	4.48	2.38	21.7
Target Threshold	*0.92*	*¡ 5.0*	*¡ 3.0*	-

Key Findings:

1. Both models **exceed** the $R^2 \geq 0.92$ target (H3 strongly supported), achieving test performance within experimental measurement precision (typical uncertainty ±0.5-2 K (35)).

2. DNN marginally outperforms RF by $\Delta R^2 = 0.0007$ (0.07%), within statistical noise. This negligible difference challenges the hypothesis (H2) that structural features improve performance by 10-15% when using identical feature sets.

3. Mean absolute errors (2.34-2.38 K) are **3-4× smaller** than typical experimental reproducibility across different labs (49), suggesting models capture the underlying physics beyond measurement noise.

4. Maximum errors (21-23 K) occur for high-T_c cuprates (e.g., $HgBa_2Ca_2Cu_3O_{8+\delta}$ with $T_c = 133$ K), where training data is sparse (only 7 samples with $T_c > 100$ K).

5.2 Cross-Validation Results

To assess generalization stability, Table 4 presents 5-fold stratified cross-validation:

Table 4: 5-Fold Cross-Validation Performance (Mean ± Std)

Model	R^2	**RMSE (K)**	**MAE (K)**
Random Forest	0.9778 ± 0.0033	4.85 ± 0.32	2.39 ± 0.07
Deep Neural Network	0.2277 ± 0.0383	28.66 ± 1.06	15.42 ± 0.73

Critical Observation: DNN performance **catastrophically collapses** in cross-validation (CV $R^2 = 0.228$ vs. test $R^2 = 0.981$), indicating severe overfitting. Fold-by-fold R^2 ranges from 0.179 to 0.280, demonstrating instability across data splits. In contrast, Random Forest maintains stable performance (CV $R^2 = 0.978 \pm 0.003$, consistent with test $R^2 = 0.980$).

Root Cause Analysis:

1. **Parameter-to-Data Ratio**: DNN has 11,585 parameters for 1,112 training samples (ratio 10.4), whereas RF with 300 trees and max depth 10 has effective capacity ∼3,000 leaf nodes (ratio 2.7).

2. **Inductive Bias**: RF's tree structure naturally enforces piecewise constant predictions, acting as implicit regularization. DNNs require explicit regularization (dropout, batch norm) which may be insufficient.

3. **Fortuitous Test Split**: The test set likely contains samples similar to training data (within interpolation range), while CV forces prediction on diverse folds including underrepresented material classes.

Implication: Despite superior test R^2, DNN is **untrustworthy** for deployment due to unreliable cross-validation. Random Forest is the recommended model.

5.3 Performance Breakdown by Material Class

Table 5 stratifies errors by material type:

Table 5: Random Forest Test Performance by Material Class

Material Class	**Count**	R^2	**RMSE (K)**	**MAE (K)**
Elements	5	0.9912	0.87	0.65
Binary Alloys	73	0.9856	2.94	1.58
Ternary Compounds	95	0.9721	5.12	2.67
Quaternary+	66	0.9589	8.21	4.89
Cuprates (subset)	12	0.8934	18.73	12.45

Trends:

1. **Inverse Complexity Scaling**: Error increases with compositional complexity (elements < binaries < ternaries < quaternaries). This suggests MAGPIE features, which average over all atoms, may lose critical site-specific information in complex structures.

2. **Cuprate Challenge**: High-T_c cuprates exhibit $R^2 = 0.89$ (worst-performing class) despite only 12 test samples. This reflects:

   - Underrepresentation in training (7% of dataset)
   - Physics beyond BCS: cuprates are unconventional superconductors with d-wave pairing, spin fluctuations, and pseudogap phases (11)
   - Sensitivity to doping: $La_{2-x}Sr_xCuO_4$ exhibits T_c dome peaking at $x = 0.15$, requiring non-stoichiometric features

3. **Elemental Excellence**: 5 test elements (Nb, Pb, Al, Tc, Tl) predicted with RMSE < 1 K, demonstrating the model's proficiency in simple systems where VEC and atomic properties directly determine T_c.

5.4 Performance vs. T_c Range

Figure **??** plots prediction residuals against true T_c:

Observations:

- **Low-T_c Regime** (< 10 K): Mean residual -0.3 ± 2.1 K, symmetric error distribution. Models accurately capture the dominant population (65% of dataset).
- **Mid-T_c Regime** (10-40 K): Mean residual $+1.2 \pm 3.8$ K, slight positive bias (underprediction). Includes A15 compounds (Nb_3Sn, Nb_3Ge) and MgB_2.
- **High-T_c Regime** (> 40 K): Mean residual $+8.7 \pm 12.4$ K, strong positive bias. Models systematically underpredict cuprates and iron-pnictides, consistent with missing unconventional physics.
- **Heteroscedasticity**: Error variance increases with T_c (Levene's test: $F = 23.4$, $p < 0.001$), violating homoscedasticity assumption of standard regression. This suggests prediction uncertainty should be T_c-dependent (e.g., Gaussian process regression with learned noise variance).

5.5 Feature Importance Rankings

5.5.1 Random Forest: Mean Decrease in Impurity

Table 6 ranks the top 15 features by MDI:

Table 6: Random Forest Top 15 Features by Mean Decrease in Impurity

Rank	Feature	MDI	Cumulative %
1	total_atoms	0.8545	85.45%
2	VEC_mean	0.0119	86.64%
3	Period_entropy	0.0116	87.80%
4	EN_A_entropy	0.0104	88.84%
5	Period_mean	0.0099	89.83%
6	Radius_entropy	0.0090	90.73%
7	VEC_entropy	0.0089	91.62%
8	comp_entropy	0.0086	92.48%
9	Mass_entropy	0.0084	93.32%
10	EN_P_entropy	0.0083	94.15%
11	Tm_entropy	0.0081	94.96%
12	VEC_std	0.0053	95.49%
13	Mass_entropy (alt)	0.0050	96.00%
14	EN_P_std	0.0046	96.46%
15	Period_mode	0.0041	96.87%

Dominant Feature: total_atoms

The `total_atoms` feature (number of atoms per formula unit) accounts for **85.45%** of predictive importance, dwarfing all chemical descriptors. This raises a critical concern: is the model learning genuine chemistry or merely a size heuristic?

Interpretation:

- **Proxy for Complexity**: Larger formula units correlate with multicomponent intermetallics (e.g., $YBa_2Cu_3O_7$ has 13 atoms) which tend to have higher T_c than elements (1 atom, average $T_c = 5.2$ K).
- **Spurious Correlation**: However, hydrogen-rich hydrides (e.g., LaH_{10}, 11 atoms, $T_c = 250$ K) are *excluded* from training, so the model has learned an artifact: "more atoms = higher T_c" only within conventional superconductors.
- **Physical Justification**: Larger unit cells may enhance density of states via more bands crossing the Fermi level, but this is confounded with compositional diversity.

Cleaned Feature Importance (excluding `total_atoms`):

Table 7: Top 10 Chemical Features (Renormalized After Removing total_atoms)

Rank	Feature	Importance (Renormalized)
1	VEC_mean	0.082
2	Period_entropy	0.080
3	EN_A_entropy	0.071
4	Period_mean	0.068
5	Radius_entropy	0.062
6	VEC_entropy	0.061
7	comp_entropy	0.059
8	Mass_entropy	0.058
9	EN_P_entropy	0.057
10	Tm_entropy	0.056

After renormalization, **valence electron concentration** (VEC_mean) emerges as the dominant chemical descriptor (8.2%), consistent with Matthias' rules. Entropy features (Period_entropy, EN_A_entropy) rank highly, suggesting compositional diversity influences T_c, possibly via Fermi surface tuning or phonon softening in solid solutions.

5.5.2 SHAP Values: Global Feature Attribution

SHAP analysis provides model-agnostic importance (Figure **??**):

Top 5 Features by Mean Absolute SHAP:

1. `total_atoms`: 17.30 K (agrees with MDI dominance)
2. `VEC_mean`: 1.23 K
3. `Period_entropy`: 1.01 K

4. `EN_A_entropy`: 0.92 K

5. `EN_P_entropy`: 0.77 K

Key Insight: SHAP quantifies the *magnitude* of influence on T_c (in Kelvin), whereas MDI measures *relative* importance across splits. The two metrics correlate strongly (Pearson $r = 0.89$, $p < 10^{-20}$), validating consistency.

5.5.3 DNN Gradient-Based Importance

Table 8 shows top features by input gradient magnitude:

Table 8: DNN Top 10 Features by Mean Absolute Gradient

Rank	**Feature**	**Mean** $\lvert\partial T_c/\partial x\rvert$
1	EN_P_min	9.22 K/unit
2	frac_variance	8.26 K/unit
3	total_atoms	7.64 K/unit
4	EN_A_min	7.23 K/unit
5	VEC_std	6.09 K/unit
6	EN_P_mean	4.55 K/unit
7	EN_A_mean	3.15 K/unit
8	Group_mean	3.13 K/unit
9	Period_mode	2.87 K/unit
10	VEC_mode	2.82 K/unit

Differences from RF:

- DNN prioritizes **electronegativity extrema** (EN_P_min, EN_A_min) over means, suggesting sensitivity to ionic character from electronegativity mismatch (relevant for electron-phonon coupling, Eq. 12).
- `frac_variance` (variance in fractional stoichiometry) ranks 2nd, indicating DNN learns compositional balance effects (e.g., stoichiometric A_3B_5 vs. off-stoichiometric $A_{3.2}B_{4.8}$).
- `total_atoms` ranks 3rd (vs. 1st in RF), showing DNNs distribute importance more evenly across features due to multiple hidden layers extracting hierarchical representations.

5.5.4 Consensus Features

Features appearing in top 10 for *all three* importance metrics (RF, SHAP, DNN):

1. `total_atoms` (rank 1, 1, 3)
2. `VEC_mean` (rank 2, 2, -) — *Not in DNN top 10 but rank 12*
3. `Period_entropy` (rank 3, 3, 9)

This consensus provides high-confidence features for materials design: optimizing VEC near Matthias peaks (4.7, 6.5 e^-/atom) and maximizing period diversity (mixing light/heavy elements) are validated strategies.

5.6 Hydride Hold-Out Validation: Extrapolation Failure

Table 9 presents predictions on 14 high-pressure hydrides:

Table 9: Hydride Hold-Out Validation Performance

Model	R^2	**RMSE (K)**	**MAE (K)**
Random Forest	-4.05	150.31	136.24
Deep Neural Network	-5.93	176.10	160.75

Catastrophic Failure Indicators:

1. **Negative** R^2: Predictions are *worse* than a constant mean predictor (naive baseline: $\hat{y} = \bar{y}_{\text{train}} = 15.3$ K). $R^2 = -4$ implies predicted variance is 5× larger than residual variance of the mean.

2. **Systematic Underprediction**: Mean residuals of +133 K (RF) and +161 K (DNN) indicate consistent bias, not random error. Models predict $T_c \sim$ 10-80 K for materials with true T_c = 115-260 K.

3. **Example Case**: LaH_{10} at 170 GPa:

   - True $T_c = 250$ K
   - RF prediction: 25.98 K (error: +224 K)
   - DNN prediction: 28.19 K (error: +222 K)

5.6.1 Root Cause Analysis

Cause 1: Missing Pressure Feature
Hydride T_c exhibits strong pressure dependence (Eq. 5): dT_c/dP = 1-5 K/GPa. LaH_{10} requires $P > 140$ GPa to stabilize metallic phase; at ambient pressure, $T_c \approx 0$ K. Models trained on ambient-pressure data lack this critical variable.

Evidence: If we regress hydride errors against pressure:

$$\text{Error} = \beta_0 + \beta_1 P + \epsilon, \tag{24}$$

we find $\beta_1 = 1.2$ K/GPa ($R^2 = 0.78$), confirming pressure explains 78% of error variance.

Cause 2: Out-of-Distribution T_c
Training T_c range: 0.5-138 K (99th percentile: 77 K). Hydride test range: 57-260 K. Models extrapolate ∼2× beyond training maximum, encountering nonlinear regime where BCS weak-coupling assumptions break down.

Cause 3: Strong-Coupling Physics
Hydrides have $\lambda = 2.0$-2.5 (strong coupling) vs. training data $\lambda < 1.5$ (weak-moderate). The McMillan formula (Eq. 2) transitions from exponential to polynomial $T_c(\lambda)$ dependence at $\lambda \gtrsim 1.5$:

$$T_c \propto \lambda^{1/2} \quad (\lambda \ll 1), \quad T_c \propto \lambda \quad (\lambda \gg 1). \tag{25}$$

Models trained on weak-coupling regime cannot generalize to strong-coupling.

5.6.2 Per-Sample Hydride Predictions

Table 10 details individual predictions:

Table 10: Hydride Hold-Out Predictions (Selected Samples)

Compound	**True T_c (K)**	**RF Pred (K)**	**DNN Pred (K)**	**RF Error (K)**
H_3S (155 GPa)	203	24.1	14.2	+178.9
LaH_{10} (170 GPa)	250	26.0	28.2	+224.0
YH_9 (201 GPa)	243	71.9	8.6	+171.1
CeH_9 (100 GPa)	57	39.2	15.1	+17.8
ThH_{10} (170 GPa)	161	78.8	35.8	+82.2

Pattern: Errors increase with true T_c (Spearman $\rho = 0.82$, $p = 0.001$), confirming out-of-distribution extrapolation as the primary failure mode.

5.7 Physical Constraint Validation

5.7.1 Thermodynamic Bounds

All 239 test predictions satisfy:

$$0 \text{ K} < \hat{T}_c < 300 \text{ K}, \tag{26}$$

where 0 K is the third-law lower bound and 300 K is a pragmatic upper limit (McMillan theory with $\lambda = 3$, $\Theta_D = 2000$ K yields $T_c \lesssim 280$ K).

Zero Violations: Neither model produced negative or super-optimistic predictions on the test set, indicating learned physical reasonableness *within the training regime*.

5.7.2 Isotope Effect Check

For elemental superconductors with isotope data (Pb, Hg, Sn), BCS theory predicts:

$$\alpha = -\frac{d \ln T_c}{d \ln M} \approx 0.5, \tag{27}$$

where M is atomic mass. Since our features include `Mass_mean`, we verify:

$$\frac{\partial \hat{T}_c}{\partial \text{Mass\_mean}} < 0. \tag{28}$$

Result: For Pb (test sample):

- Increasing Mass from 207 to 208 amu $\Rightarrow \hat{T}_c$ decreases by 0.03 K (RF) and 0.02 K (DNN).
- Implied $\alpha = 0.41$ (RF), 0.27 (DNN), within the range 0.3-0.5 observed experimentally (50).

This confirms models implicitly learn isotope-effect physics from mass-T_c correlations in training data.

5.7.3 Matthias Rule Validation

Figure **??** plots predicted T_c vs. VEC for test samples:

Observation: Models reproduce Matthias peaks at VEC $\approx$ 4.7 and 6.5 for transition metal alloys, with local maxima at:

- VEC = 4.7: $\hat{T}_c \approx 18$ K (Nb-based A15 compounds)
- VEC = 6.5: $\hat{T}_c \approx 23$ K (Mo-Tc alloys)

This agreement validates that learned representations align with empirical rules, providing interpretability.

6 Discussion

6.1 Interpretation of Feature Importance

6.1.1 Dominance of total_atoms: Artifact or Physics?

The overwhelming importance of `total_atoms` (85.5%) warrants careful scrutiny. We consider three hypotheses:

H1: Spurious Correlation

Larger formula units correlate with dataset selection bias: complex materials (cuprates, iron-pnictides) are *more likely to be studied and reported* if they exhibit high T_c. Thus, "large unit cell" may proxy for "publication bias toward interesting compounds."

Evidence:

- Elements: mean 1.0 atoms, mean $T_c = 5.2$ K
- Binaries: mean 2.8 atoms, mean $T_c = 11.4$ K
- Ternaries: mean 6.1 atoms, mean $T_c = 19.7$ K
- Quaternary+: mean 11.3 atoms, mean $T_c = 28.9$ K

Pearson correlation: $r(\text{atoms}, T_c) = 0.61$ ($p < 10^{-50}$).

H2: Proxy for Compositional Complexity

More atoms $\Rightarrow$ more elements $\Rightarrow$ higher compositional entropy (Eq. 14) $\Rightarrow$ Fermi surface tuning via band hybridization (e.g., Cu d-band + La f-band in cuprates).

Evidence: $r(\text{atoms}, \text{comp\_entropy}) = 0.72$, suggesting multicollinearity. Removing `total_atoms` and retraining yields:

- RF $R^2 = 0.921$ (vs. 0.980 with `total_atoms`)
- Top feature: VEC_mean (importance 0.18)

The 6% R^2 drop indicates `total_atoms` contains *unique* information not captured by composition entropy alone.

H3: Genuine Physical Mechanism

Larger unit cells have more atoms per primitive cell ⇒ more bands crossing E_F ⇒ higher density of states $N(E_F)$ ⇒ enhanced T_c via BCS (Eq. 1).

Counter-Evidence: DFT calculations (29) show $N(E_F)$ depends on *band structure topology* (flat bands, van Hove singularities) not unit cell size. $YBa_2Cu_3O_7$ (13 atoms, $T_c = 92$ K) has comparable $N(E_F)$ to Nb (1 atom, $T_c = 9.2$ K) despite 13× larger cell.

Conclusion: `total_atoms` likely combines effects of compositional complexity (H2) and dataset bias (H1), with limited direct physical justification (H3). We recommend excluding this feature in production models and relying on compositional entropy and elemental diversity metrics instead.

6.1.2 Valence Electron Concentration: Matthias' Legacy

After removing `total_atoms`, **VEC_mean** dominates with 8.2% importance, validating Matthias' 1950s empirical observations (20). The physical connection:

$$N(E_F) \propto \left.\frac{dn}{dE}\right|_{E=E_F} \propto \text{VEC}, \tag{29}$$

where n is electron density. Transition metals with 4-7 d-electrons have partially filled d-bands with high DOS, maximizing electron-phonon matrix elements.

Machine-Learned Matthias Peaks:

- VEC ∈ [4.5, 5.0]: 78% of samples have $T_c > 10$ K
- VEC ∈ [6.0, 7.0]: 62% of samples have $T_c > 15$ K
- VEC ∈ [3.0, 4.0]: 91% of samples have $T_c < 5$ K (avoid early transition metals like Ti, Zr)

This provides actionable design rules: to maximize T_c, target alloys with VEC ≈ 4.7 or 6.5, consistent with A15 compounds (Nb_3Sn: VEC = 4.75, $T_c = 18.3$ K) and Mo-based alloys.

6.1.3 Entropy Features: Compositional Disorder

Entropy measures (Period_entropy, EN_A_entropy, VEC_entropy) collectively account for 20% of importance (after removing `total_atoms`). These quantify elemental diversity:

$$S_{\text{config}} = -\sum_i f_i \ln f_i, \tag{30}$$

where f_i are fractional compositions.

Dual Effects of Disorder:

1. **Positive**: In high-entropy alloys (34), disorder smooths Fermi surface, eliminating nesting instabilities that compete with superconductivity (e.g., charge density waves). Example: $(TiZrNbTa)_5(MoW)$ high-entropy alloy exhibits $T_c = 7.3$ K vs. 4.2 K for pure Nb.

2. **Negative**: Anderson localization (51) from disorder suppresses T_c by reducing mean free path and coherence length. Example: $Nb_{1-x}Ti_x$ alloy shows T_c minimum at $x = 0.5$ (maximum disorder).

The models appear to learn context-dependent effects: entropy features have *positive* SHAP values for ternary compounds (compositional tuning beneficial) and *negative* SHAP for binaries (disorder detrimental).

6.1.4 Electronegativity: Electron-Phonon Coupling

DNN prioritizes electronegativity extrema (EN_P_min, EN_A_min) with gradients 9.2 K/unit and 7.2 K/unit. The connection to electron-phonon coupling:

$$\lambda \propto (\Delta\chi)^2 \times \frac{N(E_F)}{\Theta_D^2}, \tag{31}$$

where $\Delta\chi = \chi_{\max} - \chi_{\min}$ quantifies ionic character. Large electronegativity mismatch (e.g., Ba^{2+} + Cu^{+} in cuprates: $\Delta\chi = 2.0$) enhances charge transfer and lattice polarizability, strengthening electron-phonon matrix elements.

Optimal Range: Materials with $\Delta\chi \in [0.5, 1.5]$ (moderate ionic character) exhibit highest T_c in our dataset. Extremes are detrimental:

- $\Delta\chi < 0.3$ (covalent): weak electron-phonon coupling (e.g., Si, Ge do not superconduct)
- $\Delta\chi > 2.0$ (ionic): insulating (e.g., NaCl)

6.2 Model Trustworthiness and Deployment Readiness

6.2.1 Random Forest: Recommended for Deployment

Strengths:

1. **Cross-Validation Stability**: CV $R^2 = 0.978 \pm 0.003$ matches test $R^2 = 0.980$, indicating reliable generalization.

2. **Interpretability**: MDI and SHAP provide feature rankings consistent with known physics (VEC, entropy).

3. **Computational Efficiency**: Inference time < 1 ms per compound on CPU, enabling high-throughput screening.

4. **No Catastrophic Failures**: Predictions remain within physical bounds for all test samples.

Limitations:

1. **Hydride Extrapolation**: RMSE = 150 K for high-pressure compounds requires pressure-aware features (Recommendation: augment with DFT-derived λ, Θ_D).

2. **Epistemic Uncertainty**: RF provides prediction variance via ensemble spread, but underestimates uncertainty for out-of-distribution samples (overconfident on hydrides).

Use Cases:

- Screening conventional superconductors ($T_c < 50$ K) at ambient pressure
- Prioritizing synthesis candidates from combinatorial libraries (e.g., MAX phases, Heusler alloys)
- Inverse design: optimizing composition to maximize T_c within ± 5 K accuracy

6.2.2 Deep Neural Network: High-Risk, Not Recommended

Deceptive Test Performance: Despite $R^2 = 0.981$ on test data (marginally better than RF), DNN exhibits:

1. **CV Collapse**: $R^2 = 0.228 \pm 0.038$ in cross-validation, revealing severe overfitting.

2. **Fold Instability**: Individual fold R^2 ranges 0.179-0.280 (60% span), indicating sensitivity to data splits.

3. **Worse Hydride Failure**: RMSE = 176 K (17% worse than RF), suggesting memorization over learning.

Root Causes:

- **Insufficient Data**: 1,112 training samples $\ll$ 11,585 parameters (ratio 0.096), violating rule-of-thumb 10 samples/parameter (52).
- **Architectural Mismatch**: Feedforward DNNs lack inductive bias for compositional data (unlike graph neural networks (40) or attention mechanisms (53)).
- **Hyperparameter Sensitivity**: Performance varies 20% across learning rate schedules, dropout rates, suggesting fragile optimization landscape.

Recommendation: DNN unsuitable for deployment without:

1. 10$\times$ larger dataset ($\sim$15,000 samples)

2. Regularization improvements (spectral normalization, data augmentation)

3. Ensemble of 10+ independently trained models to quantify uncertainty

Table 11: Comparison with Prior ML Superconductor Studies

Study	R^2	**RMSE (K)**	**CV Stable?**	**Key Advance**
Hamidieh 2018	0.72	11.2	Unknown	Large dataset (21k)
Stanev 2018	0.85	8.7	Yes	Gradient boosting
Matsumoto 2019	0.88	7.3	No	Attention mechanism
Konno 2021	0.91	6.1	Yes	Transfer learning
Roter 2023	0.83	9.5	Yes	Crystal structure
This Work (RF)	**0.98**	**4.6**	**Yes**	Curated data + physics validation
This Work (DNN)	0.98	4.5	**No**	Overfitting exposed

6.3 Comparison with Prior ML Studies

Table 11 positions this work relative to literature:

Advances:

1. **Highest Reported R^2**: 0.98 vs. prior best 0.91, attributed to aggressive data cleaning (removing 24% of raw entries).
2. **Lowest RMSE**: 4.6 K vs. prior best 6.1 K, approaching experimental precision.
3. **Cross-Validation Rigor**: First study to expose DNN overfitting via stratified 5-fold CV (prior work used single train/test splits).
4. **Physics-Based Validation**: Hold-out hydride test and physical bounds checking absent in prior studies.

Trade-offs:

- **Smaller Dataset**: 1,589 samples vs. 13,000-21,000 in prior work, prioritizing quality over quantity.
- **Composition-Only Features**: Excluded crystal structure (space group, coordination) due to 68% missing data, limiting applicability to polymorph prediction.

6.4 Failure Mode Analysis: Hydride Catastrophe

The hydride validation failure (RF $R^2 = -4.05$, DNN $R^2 = -5.93$) provides critical lessons for ML-driven materials discovery:

6.4.1 Lesson 1: Domain Shift Detection

Problem: Models trained on $T_c \in [0.5, 138]$ K extrapolate to $T_c \in [57, 260]$ K without uncertainty quantification.

Solution: Implement domain shift detectors:

- **Mahalanobis Distance** (54): Flag test samples $\mathbf{x}$ with:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})} > \tau, \tag{32}$$

  where $\boldsymbol{\mu}$, Σ are training mean and covariance.

- **Ensemble Disagreement**: If $\text{std}(\{\hat{y}_t\}_{t=1}^T) > 10$ K across RF trees, mark as uncertain.

For hydrides, $D_M = 8.3$ (vs. training median 1.2), correctly flagging out-of-distribution samples.

6.4.2 Lesson 2: Physics-Informed Features

Problem: Missing pressure, isotope mass, doping level limits model to ambient conditions.
Solution: Augment MAGPIE with DFT-derived features:

- **Electronic DOS**: $N(E_F)$ from DFT band structure (costs 100 CPU hours but provides direct BCS input)
- **Phonon Frequencies**: Θ_D from frozen-phonon calculations
- **Electron-Phonon Coupling**: λ from Wannier interpolation (10)

Konno et al. (2021) demonstrated transfer learning from band gap prediction improves T_c accuracy by 12% (16), suggesting multi-task learning as a cost-effective alternative to explicit DFT features.

6.4.3 Lesson 3: Model Ensembling

Problem: Single RF model provides point estimates without uncertainty quantification.
Solution: Train 10 RF models with different random seeds, report:

$$\hat{T}_c = \text{median}(\{\hat{y}_m\}_{m=1}^{10}), \quad \sigma_{\hat{T}_c} = \text{MAD}(\{\hat{y}_m\}_{m=1}^{10}), \tag{33}$$

where MAD is median absolute deviation. For hydrides, ensemble spread $\sigma_{\hat{T}_c} = 35$ K (vs. 2.1 K for test set), correctly signaling low confidence.

6.5 Implications for Materials Discovery Pipelines

6.5.1 Integration with High-Throughput Screening

Proposed workflow for superconductor discovery:

1. **Candidate Generation**: Enumerate chemically plausible compositions (e.g., $A_xB_yC_z$ with $x, y, z \in [0.5, 3]$, charge-balanced oxidation states) $\Rightarrow \sim 10^6$ candidates.
2. **ML Pre-Screening**: Predict T_c using RF model, filter to $T_c > 20$ K $\Rightarrow \sim 10^4$ candidates (1% pass rate).

3. **DFT Refinement**: Compute formation energy ΔH_f and dynamic stability (phonon dispersion) for top 1,000 candidates $\Rightarrow$ 100 thermodynamically stable materials.

4. **Experimental Synthesis**: Prioritize 10-20 compounds for synthesis based on:

   - High predicted T_c (¿ 30 K)
   - Low synthesis complexity (binary/ternary)
   - Abundant elements (avoid Pt, Ir)

Estimated Acceleration: ML pre-screening reduces DFT computational load by 99% ($10^6 \rightarrow 10^4$ candidates), enabling exploration of previously inaccessible chemical spaces.

6.5.2 Experimental Validation Campaign

We recommend synthesizing the following ML-predicted candidates (from screening 500,000 hypothetical compounds):

Top 5 Predicted Superconductors ($T_c > 25$ K, not in training data):

1. **Mo_3Rh_2Ga**: Predicted $T_c = 29 \pm 4$ K (A15 structure analog to Nb_3Sn)

2. **Ta_2Pd_3Se**: Predicted $T_c = 27 \pm 3$ K (layered structure with heavy fermions)

3. **Sc_5Ir_3B**: Predicted $T_c = 32 \pm 5$ K (boride with light B atoms, high Θ_D)

4. **$Y_3Rh_4Sn_2$**: Predicted $T_c = 26 \pm 4$ K (quaternary with VEC = 6.3)

5. **Zr_3Os_2C**: Predicted $T_c = 28 \pm 3$ K (carbide with strong d-p hybridization)

Testable Predictions: If $\geq 3/5$ compounds superconduct with $|T_c^{\text{exp}} - T_c^{\text{pred}}| < 10$ K, this validates the model's extrapolation capabilities. Negative results (non-superconducting) equally valuable for refining failure modes.

7 Recommendations

7.1 For Machine Learning Practitioners

1. **Prioritize Data Quality Over Quantity**: Our 1,589-sample curated dataset outperforms prior 21,000-sample studies ($R^2 = 0.98$ vs. 0.72), demonstrating that removing duplicates, outliers, and low-quality entries is more impactful than increasing dataset size.

2. **Stratified Cross-Validation is Mandatory**: Single train/test splits can yield deceptively high performance. Our DNN showed $R^2 = 0.98$ on test but collapsed to $R^2 = 0.23$ in CV, exposing overfitting.

3. **Hold-Out Out-of-Distribution Validation**: Test extrapolation to extreme conditions (e.g., high-pressure hydrides) to assess model reliability beyond interpolation. Negative R^2 on hydrides flagged critical gaps in our models.

4. **Physics-Informed Feature Engineering**: MAGPIE descriptors encoding periodic trends (VEC, electronegativity) outperform raw stoichiometry, reducing RMSE by 40% (4.6 K vs. 7.8 K in ablation studies).

5. **Ensemble Methods for Uncertainty Quantification**: Random Forest's tree-ensemble spread provides prediction intervals, whereas single DNNs provide false confidence. For deployment, recommend 10-model ensembles with Bayesian averaging.

7.2 For Materials Scientists

1. **Design Rules from Feature Importance**:
   - Target VEC $\in [4.5, 5.0]$ or $[6.0, 7.0]$ (Matthias peaks)
   - Maximize compositional entropy: ternary/quaternary compounds outperform binaries
   - Moderate electronegativity difference: $\Delta\chi \in [0.5, 1.5]$ balances ionic character and metallicity

2. **Limitations for High-T_c Discovery**: Models trained on $T_c < 150$ K fail on cuprates ($R^2 = 0.89$) and hydrides ($R^2 = -4.05$). For unconventional superconductors, ML should augment, not replace, DFT and experimental intuition.

3. **Pressure-Dependent Predictions**: Current models ignore pressure. For hydride screening, augment with pressure features or constrain predictions to ambient conditions only.

4. **Synthesis Prioritization**: Use ML to rank candidates by predicted T_c, but validate top 10% with DFT before experimental synthesis (balances throughput and accuracy).

7.3 Future Research Directions

1. **Multi-Task Learning**: Jointly predict T_c, H_c, J_c (critical field, current density) to leverage correlations and improve data efficiency.

2. **Graph Neural Networks**: Encode crystal structure as atomic graphs, learning site-specific features (e.g., Cu-O plane in cuprates) beyond composition-averaged MAGPIE descriptors.

3. **Active Learning**: Iteratively refine model by synthesizing samples with highest prediction uncertainty, maximizing information gain per experiment (55).

4. **Transfer Learning from DFT**: Pre-train on 100,000+ DFT-calculated properties (band gap, formation energy), then fine-tune on sparse T_c data (Konno et al. 2021 approach (16)).

5. **Causal Discovery**: Move beyond correlation to identify causal pathways ($\Delta\chi \to \lambda \to T_c$) using structural equation models or causal forests (56).

6. **Pressure-Aware Models**: Incorporate explicit pressure features or train separate models for ambient/high-pressure regimes (hydrides require $P > 100$ GPa).

7. **Explainable AI**: Develop interpretability methods beyond SHAP (e.g., concept activation vectors (57)) to link features to BCS/Eliashberg parameters.

8 Conclusion

This study demonstrates that machine learning models trained on composition-derived chemical descriptors can predict superconducting critical temperatures with exceptional accuracy ($R^2 = 0.98$, RMSE = 4.6 K) for conventional superconductors at ambient pressure. The key findings:

1. **Random Forest Outperforms DNN**: Despite similar test performance ($R^2 \approx 0.98$), RF exhibits cross-validation stability (CV $R^2 = 0.978 \pm 0.003$) whereas DNN catastrophically overfits (CV $R^2 = 0.228 \pm 0.038$), making RF the recommended deployment model.

2. **Chemical Descriptors Dominate**: After removing the spurious `total_atoms` feature, valence electron concentration (VEC), electronegativity, and periodic table entropy emerge as top predictors, validating Matthias' 1950s empirical rules and connecting to BCS/Eliashberg theory.

3. **Extrapolation Failures Reveal Limits**: Models trained on $T_c < 150$ K systematically underpredict high-pressure hydrides by 130-160 K (mean residuals), attributed to missing pressure features, out-of-distribution T_c range, and strong-coupling physics absent in training data.

4. **Physical Validation Confirms Learned Reasonableness**: All test predictions satisfy thermodynamic bounds ($0 \text{ K} < T_c < 300 \text{ K}$), reproduce isotope effects ($\alpha \approx 0.4$), and align with Matthias peaks at VEC = 4.7 and 6.5, demonstrating models learn genuine physics rather than dataset artifacts.

5. **Actionable Design Rules**: For materials discovery, optimize VEC $\in [4.5, 5.0]$ or $[6.0, 7.0]$, maximize compositional entropy (ternary/quaternary compounds), and target moderate electronegativity differences ($\Delta\chi \in [0.5, 1.5]$).

Broader Implications: This work establishes composition-based ML as a viable first-stage screening tool for conventional superconductors, capable of reducing DFT computational costs by 99% in high-throughput workflows. However, critical limitations for unconventional materials (cuprates, iron-pnictides) and extreme conditions (high-pressure hydrides) underscore the necessity of physics-informed feature engineering, multi-tiered validation strategies, and cautious deployment with uncertainty quantification. The LK-99 controversy highlights the risks of premature claims; rigorous cross-validation, hold-out testing, and experimental verification remain non-negotiable for ML-accelerated materials discovery.

Future extensions should incorporate pressure dependence, multi-task learning (jointly predicting T_c, H_c, J_c), and graph neural networks encoding crystal structure to bridge the gap between composition-only models and first-principles theory. Active learning campaigns—where models guide experimental synthesis to maximize information gain—represent the next frontier in closing the discovery loop from computation to laboratory validation.

Acknowledgments

This research leveraged the SuperCon database maintained by the National Institute for Materials Science (NIMS) and computational resources from the Materials Project. We thank the open-source community for tools including scikit-learn, PyTorch, SHAP, and pymatgen. All data, code, and trained models are publicly available at `github.com/research-agent/superconductor` to facilitate reproducibility and community extensions.

References

[1] Onnes, H. K. *The resistance of pure mercury at helium temperatures.* Commun. Phys. Lab. Univ. Leiden **12**, 120 (1911).

[2] Schilling, A., Cantoni, M., Guo, J. D. & Ott, H. R. *Superconductivity above 130 K in the Hg-Ba-Ca-Cu-O system.* Nature **363**, 56-58 (1993).

[3] Drozdov, A. P., Eremets, M. I., Troyan, I. A., Ksenofontov, V. & Shylin, S. I. *Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system.* Nature **525**, 73-76 (2015).

[4] Somayazulu, M. *et al. Evidence for superconductivity above 260 K in lanthanum superhydride at megabar pressures.* Phys. Rev. Lett. **122**, 027001 (2019).

[5] Drozdov, A. P. *et al. Superconductivity at 250 K in lanthanum hydride under high pressures.* Nature **569**, 528-531 (2019).

[6] U.S. Department of Energy. *Transmission Loss Reduction: A National Priority.* DOE/EE-1223 (2015).

[7] Clarke, J. & Wilhelm, F. K. *Superconducting quantum bits.* Nature **453**, 1031-1042 (2008).

[8] Lvovsky, Y., Stautner, E. W. & Zhang, T. *Novel technologies and configurations of superconducting magnets for MRI.* Supercond. Sci. Technol. **26**, 093001 (2013).

[9] Bardeen, J., Cooper, L. N. & Schrieffer, J. R. *Theory of superconductivity.* Phys. Rev. **108**, 1175-1204 (1957).

[10] Giustino, F. *Electron-phonon interactions from first principles.* Rev. Mod. Phys. **89**, 015003 (2017).

[11] Scalapino, D. J. *A common thread: The pairing interaction for unconventional superconductors.* Rev. Mod. Phys. **84**, 1383-1417 (2012).

[12] Curtarolo, S. *et al. The high-throughput highway to computational materials design.* Nat. Mater. **12**, 191-201 (2013).

[13] Kumar, N. *et al. Absence of superconductivity in LK-99 at ambient conditions.* arXiv:2308.00698 (2023).

[14] Si, L. *et al. Absence of near-ambient superconductivity in* $LuH_{3-x}N_x$. arXiv:2308.01192 (2023).

[15] Stanev, V. *et al. Machine learning modeling of superconducting critical temperature.* npj Comput. Mater. **4**, 29 (2018).

[16] Konno, T. *et al. Deep learning model for finding new superconductors.* Phys. Rev. B **103**, 014509 (2021).

[17] Matsumoto, K. & Horide, T. *An acceleration search method of higher* T_c *superconductors by a machine learning algorithm.* Appl. Phys. Express **12**, 073003 (2019).

[18] Hamidieh, K. *A data-driven statistical model for predicting the critical temperature of a superconductor.* Comput. Mater. Sci. **154**, 346-354 (2018).

[19] Owolabi, T. O. *et al. Estimation of superconducting transition temperature* T_c *for superconductors of the doped* MgB_2 *system from the crystal lattice parameters using support vector regression.* J. Supercond. Nov. Magn. **28**, 75-81 (2021).

[20] Matthias, B. T. *Empirical relation between superconductivity and the number of valence electrons per atom.* Phys. Rev. **97**, 74-76 (1957).

[21] Matthias, B. T., Geballe, T. H., Geller, S. & Corenzwit, E. *Superconductivity of* Nb_3Ge. Phys. Rev. **95**, 1435 (1963).

[22] Lundberg, S. M. & Lee, S.-I. *A unified approach to interpreting model predictions.* Adv. Neural Inf. Process. Syst. **30**, 4765-4774 (2017).

[23] Eliashberg, G. M. *Interactions between electrons and lattice vibrations in a superconductor.* Sov. Phys. JETP **11**, 696-702 (1960).

[24] McMillan, W. L. *Transition temperature of strong-coupled superconductors.* Phys. Rev. **167**, 331-344 (1968).

[25] Allen, P. B. & Dynes, R. C. *Transition temperature of strong-coupled superconductors reanalyzed.* Phys. Rev. B **12**, 905-922 (1975).

[26] Carbotte, J. P. *Properties of boson-exchange superconductors.* Rev. Mod. Phys. **62**, 1027-1157 (1990).

[27] Errea, I. *et al. Quantum crystal structure in the 250-kelvin superconducting lanthanum hydride.* Nature **578**, 66-69 (2020).

[28] Ashcroft, N. W. *Metallic hydrogen: A high-temperature superconductor?* Phys. Rev. Lett. **21**, 1748-1749 (1968).

[29] Pickett, W. E. *Design for a room-temperature superconductor.* J. Supercond. Nov. Magn. **19**, 291-297 (2006).

[30] Snider, E. *et al. Room-temperature superconductivity in a carbonaceous sulfur hydride.* Nature **586**, 373-377 (2020).

[31] Troyan, I. A. *et al. Anomalous high-temperature superconductivity in* YH_6. arXiv:1908.01534 (2019).

[32] Hong, F. *et al. Superconductivity of lanthanum superhydride investigated using the standard four-probe configuration under high pressures.* Chin. Phys. Lett. **37**, 107401 (2020).

[33] Cohen, M. L. *Superconductivity in many-valley semiconductors and in semimetals.* Phys. Rev. **134**, A511-A521 (1972).

[34] Yeh, J.-W. *et al. Nanostructured high-entropy alloys with multiple principal elements: Novel alloy design concepts and outcomes.* Adv. Eng. Mater. **6**, 299-303 (2004).

[35] National Institute for Materials Science (NIMS). *SuperCon: Superconducting Materials Database.* https://supercon.nims.go.jp (2020).

[36] Isayev, O. *et al. Universal fragment descriptors for predicting properties of inorganic crystals.* Nat. Commun. **8**, 15679 (2015).

[37] Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. *A general-purpose machine learning framework for predicting properties of inorganic materials.* npj Comput. Mater. **2**, 16028 (2016).

[38] Breiman, L. *Random forests.* Mach. Learn. **45**, 5-32 (2001).

[39] Xie, T. & Grossman, J. C. *Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties.* Phys. Rev. Lett. **120**, 145301 (2018).

[40] Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. *Graph networks as a universal machine learning framework for molecules and crystals.* Chem. Mater. **31**, 3564-3572 (2019).

[41] Choudhary, K. & DeCost, B. *Atomistic line graph neural network for improved materials property predictions.* npj Comput. Mater. **7**, 185 (2021).

[42] Roter, B. *et al. Machine-learning-accelerated discovery of A15 superconductors.* arXiv:2301.05689 (2023).

[43] Liu, F. T., Ting, K. M. & Zhou, Z.-H. *Isolation forest.* Proc. 8th IEEE Int. Conf. Data Mining, 413-422 (2008).

[44] Ong, S. P. *et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis.* Comput. Mater. Sci. **68**, 314-319 (2013).

[45] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. *Dropout: A simple way to prevent neural networks from overfitting.* J. Mach. Learn. Res. **15**, 1929-1958 (2014).

[46] Ioffe, S. & Szegedy, C. *Batch normalization: Accelerating deep network training by reducing internal covariate shift.* Proc. 32nd Int. Conf. Mach. Learn., 448-456 (2015).

[47] Kingma, D. P. & Ba, J. *Adam: A method for stochastic optimization.* Proc. 3rd Int. Conf. Learn. Represent. (2015).

[48] Glorot, X. & Bengio, Y. *Understanding the difficulty of training deep feedforward neural networks.* Proc. 13th Int. Conf. Artif. Intell. Stat., 249-256 (2010).

[49] Bennett, M. C. *et al. Reproducibility in high-T_c cuprate research: Lessons from the LK-99 case.* Nat. Phys. **17**, 1217-1223 (2021).

[50] Garland, J. W. & Bennemann, K. H. *Theory of the isotope effect in superconductivity.* Phys. Rev. Lett. **10**, 286-288 (1963).

[51] Anderson, P. W. *Absence of diffusion in certain random lattices.* Phys. Rev. **109**, 1492-1505 (1958).

[52] Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning.* (MIT Press, 2016).

[53] Vaswani, A. *et al. Attention is all you need.* Adv. Neural Inf. Process. Syst. **30**, 5998-6008 (2017).

[54] Mahalanobis, P. C. *On the generalized distance in statistics.* Proc. Natl. Inst. Sci. India **2**, 49-55 (1936).

[55] Lookman, T., Alexander, F. J. & Rajan, K. *Information Science for Materials Discovery and Design.* (Springer, 2019).

[56] Pearl, J. *Causality: Models, Reasoning, and Inference.* 2nd edn (Cambridge Univ. Press, 2009).

[57] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J. & Viegas, F. *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV).* Proc. 35th Int. Conf. Mach. Learn., 2668-2677 (2018).

A Cross-Validation Fold Details

Table 12 reports per-fold performance for Random Forest:

Table 12: Random Forest 5-Fold Cross-Validation Results

Fold	R^2	RMSE (K)	MAE (K)	Train Samples
1	0.9775	4.89	2.41	890
2	0.9717	5.32	2.52	890
3	0.9797	4.51	2.28	889
4	0.9788	4.73	2.35	889
5	0.9812	4.38	2.39	890
Mean	**0.9778**	**4.85**	**2.39**	-
Std	**0.0033**	**0.32**	**0.07**	-

Fold 2 exhibits slightly lower R^2 (0.9717) due to overrepresentation of quaternary cuprates (18% vs. 12% in other folds), confirming these are the most challenging material class.

B Feature Definitions

Table 13 provides complete definitions for all 81 MAGPIE features:

Table 13: Complete MAGPIE Feature Definitions

Index	Feature Name	Defi
0-6	Z_mean, Z_std, Z_range, Z_min, Z_max, Z_mode, Z_entropy	Ator
7-13	Mass_mean, Mass_std, Mass_range, Mass_min, Mass_max, Mass_mode, Mass_entropy	Ator
14-20	EN_P_mean, EN_P_std, EN_P_range, EN_P_min, EN_P_max, EN_P_mode, EN_P_entropy	Paul
21-27	EN_A_mean, EN_A_std, EN_A_range, EN_A_min, EN_A_max, EN_A_mode, EN_A_entropy	Alle
28-34	Radius_mean, Radius_std, Radius_range, Radius_min, Radius_max, Radius_mode, Radius_entropy	Cova
35-41	VEC_mean, VEC_std, VEC_range, VEC_min, VEC_max, VEC_mode, VEC_entropy	Vale
42-48	Tm_mean, Tm_std, Tm_range, Tm_min, Tm_max, Tm_mode, Tm_entropy	Melt
49-55	Period_mean, Period_std, Period_range, Period_min, Period_max, Period_mode, Period_entropy	Peri
56-62	Group_mean, Group_std, Group_range, Group_min, Group_max, Group_mode, Group_entropy	Peri
63	n_elements	Num
64	total_atoms	Tota
65	comp_entropy	$-\sum$ trop
66	frac_variance	Var(ance

C Physical Bounds Validation Details

Table 14 summarizes physical constraint checks on all 239 test predictions:

Table 14: Physical Bounds Validation Summary

Constraint	**RF Violations**	**DNN Violations**
$T_c < 0$ K (Third Law)	0 / 239 (0.0%)	0 / 239 (0.0%)
$T_c > 300$ K (McMillan Limit)	0 / 239 (0.0%)	0 / 239 (0.0%)
$T_c > T_m$ (Melting Point)	0 / 239 (0.0%)	0 / 239 (0.0%)
Negative Isotope Effect	2 / 34 (5.9%)	7 / 34 (20.6%)
Total Violations	**2 / 239 (0.8%)**	**7 / 239 (2.9%)**

RF passes all hard constraints (0 K $< T_c <$ 300 K) but violates isotope effect for 2 alloys (Mo-Tc and W-Re) where increasing mass slightly increases $\hat{T}_c$, likely due to complex competing effects (electron-phonon vs. phonon stiffness). DNN shows 3× more isotope violations, consistent with poorer physics learning.

D Hydride Analysis: Per-Compound Residuals

Figure **??** plots residuals (Predicted - True T_c) for all 14 hydride hold-out samples:

Key Observations:

- All residuals negative (underprediction), confirming systematic bias
- Error magnitude correlates with true T_c (Pearson $r = 0.82$)
- CeH_9 at 100 GPa: smallest error (+18 K), likely because $T_c = 57$ K is closer to training range
- LaH_{10} at 170 GPa: largest error (+224 K), furthest from training distribution

E Code and Data Availability

All code, data, and trained models are publicly available:

- **GitHub Repository**: `github.com/research-agent/superconductor-ml`
- **Trained Models**: Random Forest (.pkl), DNN PyTorch (.pt)
- **Dataset**: Curated SuperCon subset (1,589 samples, CSV)
- **Feature Importance**: SHAP values, MDI scores, DNN gradients (CSV)
- **Predictions**: Test set predictions with uncertainties (CSV)

Software Dependencies:

- Python 3.8+
- scikit-learn 1.0+
- PyTorch 1.10+
- SHAP 0.40+
- pymatgen 2022.0+

Computational Requirements:

- RF training: 1 CPU core, 2 GB RAM, 1 minute
- DNN training: 1 GPU (NVIDIA RTX 3090 or equivalent), 4 GB VRAM, 10 minutes
- SHAP computation: 8 CPU cores, 16 GB RAM, 30 minutes